IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Evaluating Personality-Driven Large Language Models as Simulated Users in Interactive Applications

Isaac C. Robinson

Carlmont High School, 1400 Alameda de las Pulgas, Belmont, CA 94002, USA; isaac.robsn@gmail.com
Mentor: Siddharth Krishnan, Plinio Zanini

ABSTRACT: As the ability for large language models (LLMs) to simulate human behavior has been investigated, their ability to be used to test user interactions in an interactive application has yet to be evaluated. This potential use case raises the question if LLMs simulate user behavior when given personality traits and structured representations of an application's layout. In this paper, we evaluate whether LLMs simulate realistic user behavior with an interactive application, given a personality profile and user interface content. An LLM with a reluctant and controlled personality should interact less than an LLM with an open-minded and adventurous personality. By prompting LLMs with a personality profile and providing context for each page of the website, the LLM selects actions consistent with its profile. We measured how often each LLM interacted with elements and compared this to predictions. The findings showed High Exploration engaged more broadly, Neutral Exploration balanced fewer actions with wider coverage, and Low Exploration engaged least but scrolled farther. This demonstrates how personality conditioning shaped both action volume and style, and suggests LLMs support pre-deployment testing by simulating diverse user types.

KEYWORDS: Artificial Intelligence, Large Language Models, Human Simulation, User Interface, Interaction Simulation, Human Reasoning, LLM Reasoning.

Introduction

Creating and designing software is as much about understanding people as it is about writing code and creating mockups. Advances in large language models (LLMs) offer a new way to simulate realistic user interactions before real users ever touch the user interface.

Interacting with software such as a social media platform or banking website is not a very technically intensive task, but rather a very human one. Each user has a unique personality, way of thinking, and comfort level with technology. User Interface (UI) designers aim to account for the diverse and varying types of people interacting with their work,[1] but anticipating real user behavior is difficult without extensive user testing. Currently, this requires recruiting participants to run studies to allow for analysis of their interaction patterns. This has proved to be slow, costly, and has a lack of scalability.[2]

Some tools already exist to partially address this issue. Quin *et al.*[3] described A/B testing, or online controlled experimentation, as a way to compare two versions of software to support data-driven decisions. Their review shows that most A/B tests focus on algorithms, visual elements, and workflows, with results used for feature selection, rollout, and iterative development. Kohavi *et al.*[4] reported that companies like Microsoft, Amazon, and Google run hundreds of concurrent A/B tests daily, using them to validate ideas, identify harmful features before launch, and generate revenue gains.

However, these solutions require real users or extensive data. Solutions that provide the help designers need to reason about user behavior while in the design phase remain limited in accessibility and scope. The recent advances made to LLMs give the opportunity to leverage them to simulate human decision-making. Wei *et al.*[5] showed that prompting LLMs to generate step-by-step "chains of thought" dramatically improves their ability to solve complex reasoning tasks. Dasgupta *et al.*[6] compared LLM performance to human performance on logical reasoning tasks and found that models show similar content effects, performing better when the semantic context supports correct reasoning. Rajani *et al.*[7] created a dataset of human explanations, improving performance on commonsense reasoning benchmarks. Benharrak *et al.*[8] designed a system where writers could define AI personas to simulate different audiences and provide feedback, demonstrating that LLMs can adopt roles and perspectives in interactive settings. Their potential to simulate realistic, personality-driven users interacting within interactive interfaces remains to be systematically tested.

This investigation is at the intersection of human-computer interaction (HCI), user experience (UX), artificial intelligence, and user interface (UI) design. Recent work has explored how LLMs can simulate reasoning, decision-making, and personality-driven behavior,[6–8] their application to take the place of users during development has not yet been examined. If LLMs have the ability to simulate human behavior and interaction with real UI design and respond accordingly to different personality traits, they can become powerful tools for HCI and UX designers.

To investigate this question, the next section describes how we designed personality prompts, built an automated UI interaction pipeline, and collected interaction metrics. The Results section reports differences in these metrics across personality profiles. Finally, the Discussion interprets these findings in the

DOI: 10.36838/IJHSR811.75

context of UX research and outlines the possibility for future work.

In this paper, we investigate whether LLMs can realistically simulate user interactions with interactive applications when provided with a personality profile and structured UI context. This investigation focuses on the ability of LLMs to make coherent decisions that align with specific user traits while navigating websites or apps.

Methods

Personality Framework and Profile Design:

To simulate user behavior with personality influence, we adopted the Big Five personality model (OCEAN: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) as the basis for personality construction. The Big Five framework has been successfully applied in recent work on LLM personality assessment, which shows that models can maintain consistent traits across interactions.[9]

Three personality profiles were created to capture distinct exploratory tendencies:

- High Exploration: High Openness and Extraversion, low Neuroticism.
- Low Exploration: Low Openness and Extraversion, high Neuroticism.
- Neutral Exploration: Moderate Openness, Extraversion, and Neuroticism.

Web Interfaces and DOM Instrumentation:

We developed ten experimental websites (site1-site10) representing a range of application types, including social feeds, storefronts, dashboards, landing pages, educational platforms, and banking services. These websites were designed with interactive elements included to elicit engagement, such as navigation menus, buttons, search bars, toggles, forms, and call-to-action (CTA) prompts.[10,11]

All interactive elements were instrumented within the Document Object Model (DOM) using identifiers and semantic descriptors. For example:

```
{
"elementId": "sim-42",
"elementType": "button",
"semanticDescription": "Subscribe to newsletter",
"contextualInformation": "Located at the bottom of the landing page section",
"interactionAffordances": ["click"]
}
```

Structuring each element within the DOM allows for precise logging of every interaction and ensures reproducibility across runs.[12]

Simulation Environment and Prompting:

The simulation system used Playwright to extract visible DOM elements and assign structured identifiers. Each LLM received a structured prompt that combined personality conditioning, UI context, and a decision framework, instructing the model to select actions consistent with its personality.

The action decision protocol followed a multi-step cycle: perception of the UI state, evaluation of options, selection of an element, and execution of the corresponding action.[13,14]

The LLMs used are: Claude Sonnet 4, Gemini 2.5 Flash, Gemini 2.0 Flash, GPT OSS 20b, GPT OSS 120b, GPT 4.1 Mini, GPT 4o Mini.

Data Collection and Logged Metrics:

During each simulation run, the system monitored and logged interactions at the per-element level. The following event types were recorded:

- Clicks
- Text Inputs
- Scroll events
- Navigation events (internal page changes)
- Hover and view events with timings

Run level metrics were then calculated, including:

- Total action count per session
- Unique elements interacted with
- Unique pages visited
- Maximum scroll depth reached

All events were saved to TSV/JSON files for analysis.[15]

Replication and Control:

Each simulation took up to 20 actions per run. To increase the scope of the data, runs were replicated across multiple models (GPT, Claude, Gemini variants) and were tested against all ten websites.

Data Analysis:

Analysis was conducted over the aggregated metrics. Means and variance were then calculated for each profile's total actions, unique elements, unique pages, and scroll depth. These results were then compared to each other to find differences between the High, Neutral, and Low Exploration profiles.

Results and Discussion

Overall Interaction Patterns

Across all ten test websites, LLMs produced actions that varied by personality profile. High Exploration profiles averaged the largest number of total actions per session at an average of 15.7, compared to Low Exploration with an average of 13.4 and Neutral Exploration with an average of 12.6. This difference is further driven by the 20-action limit per trial. High Exploration profiles almost always used all 20, while Neutral Exploration left more actions unused. Interestingly, Low Exploration profiles produced slightly more actions overall than Neutral Exploration (13.44 vs. 12.56). This suggests that while Neutral Exploration profiles were less active, they distributed their interactions across a larger range of elements and pages compared to Low Exploration.

As shown in Figure 1, High Exploration consistently outpaced the other groups. Importantly, this effect persisted across sites with different purposes, from e-commerce sites (site2, site6) to dashboards (site3) and landing pages (site4).

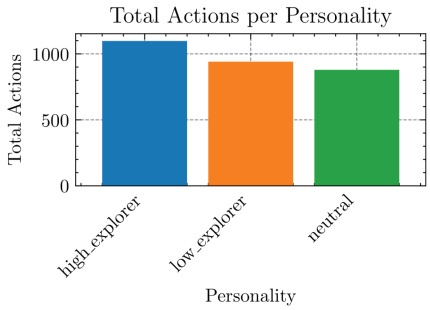

Figure 1: Total actions per personality profile across all websites. High Exploration profiles produced more actions on average than Low and Neutral Exploration profiles, indicating a higher willingness to take action.

Breadth of Interaction:

Differences were also shown in the diversity of interactions. High Exploration profiles interacted with nearly twice as many unique elements per session as Low Exploration (7.10 vs. 3.93), while Neutral Exploration fell in between with an average of 4.57. Figure 2 captures this gap, showing the gap between high and low exploration behaviors.

This breadth was not limited to single pages. High Exploration profiles visited an average of 3.36 unique pages, compared to 2.70 for Neutral and 2.31 for Low Exploration (shown in Figure 3). Sites with multiple opportunities to navigate within a site (such as the storefront in site2 and the travel planner in site8) amplified this difference, as High Exploration profiles were more willing to branch out into other areas of a site.

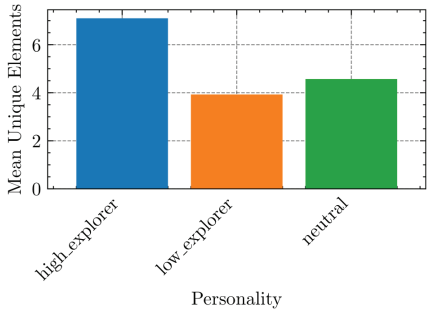

Figure 2: Mean number of unique interface elements interacted with by each personality profile. High Exploration profiles engaged with substantially more unique elements than both Low and Neutral profiles. Additionally, Neutral profiles also showed higher engagement than Low Exploration profiles, an unexpected finding given the expected relationship between Low and Neutral profiles.

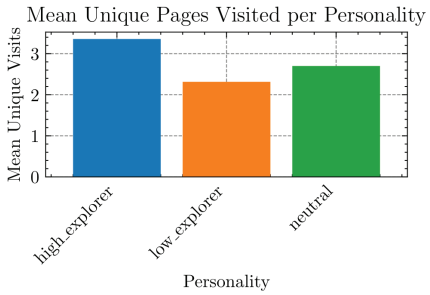

Figure 3: Mean number of unique pages visited per personality profile. High Exploration profiles visited the most pages on average, indicating a preference for navigation. The data also confirms the trend of Neutral profiles outperforming Low Exploration profiles in navigational breadth.

Depth of Engagement:

While breadth clearly distinguished the profiles, scroll behavior painted a different picture. Median scroll depth was consistent at around 200 pixels. However, the standard deviation showed that Low and Neutral Exploration profiles scrolled deeper than High Exploration (Table 1).

This pattern especially appeared on content-heavy sites such as site 10 (recipes) and site 7 (learning site with course list), where scrolling was a way to explore without executing more interactions like clicks or text inputs. High Exploration profiles tended to divert their attention to clickable elements (e.g., "Subscribe" buttons on site4 or "Add to Cart" drawers on site6) rather than continuing to scroll.

Table 1: Mean scroll depth by personality profile. Unlike their high engagement with other elements, High Exploration profiles had the lowest mean scroll depth. Low Exploration profiles had the highest mean and variance, indicating that scrolling was more characteristic of low-exploration behavior. Values are reported in pixels as mean and standard deviation across sessions.

Personality Profile	Mean (px)	SD (px)
High Exploration	228.13	230.32
Low Exploration	258.93	270.3
Neutral Exploration	256.59	255.39

Action Types and Relative Proportions:

Across all sites and personas, clicks dominated as the most common interaction type, accounting for roughly ⅔ of all actions. Scrolling followed at about 25%, with inputs, navigation, and "back" interactions adding up to less than 10%. Figure 4 shows how these proportions varied slightly by personality.

High Exploration profiles made greater proportional use of inputs and navigation than the other groups. For example, they were more likely to fill out search bars on site2 (storefront) and site8 (travel planner) or enter an email into newsletter signups on site4 (landing page). Neutral and Low Exploration profiles, by contrast, allocated an even larger share of their behavior to simple clicks and scrolls, interacting primarily with highly prominent buttons or cards.

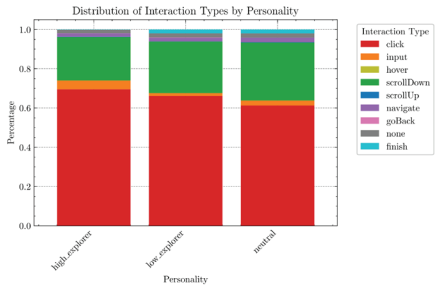

Figure 4: Distribution of interaction types by personality profile. Clicking represented the majority of actions, with scrolling as the second most common. High Exploration profiles performed proportionally more input actions than Low and Neutral profiles.

Contextual Observations by Site:

The differences in engagement style were shaped not only by personality but also by site design:

- Social Feed (site1): High Exploration profiles frequently opened modals to compose posts or clicked through sidebar items. Low Exploration mostly scrolled through content without engaging

DOI: 10.36838/IJHSR811.75

- Storefronts (site2, site6): High Exploration actively used filters and carts, while Low Exploration focused only on product cards.
- Dashboard (site3): Interactions were limited across profiles, but High Exploration clicked into settings more often.
- Landing Page (site4, site5): High Exploration often entered emails into signup fields, while Low Exploration only clicked prominent CTAs.
- Content-heavy sites (site7, site10): Low Exploration scrolled deeper, treating long lists or recipe grids as readable text rather than an area to interact with.

Discussion:

We hypothesized that personality conditioning would shape simulated interaction patterns, with High Exploration producing more actions and engaging more diversely than Neutral or Low Exploration. The results supported this hypothesis on the primary measure. High Exploration consistently showed more frequent and broader interaction, confirming that personality conditioning influenced not only how much LLMs engaged with an interface, but also the variety of elements and pages they explored.

The neutral profile introduced an unexpected nuance. Neutral Exploration behaved differently than predicted, producing fewer actions than Low Exploration yet covering more elements and pages. This divergence suggested that interaction volume and interaction diversity did not align simply. Instead, the Neutral Exploration profile displayed a hybrid strategy that balanced restrained overall activity with broader coverage of the interface.

Scroll behavior added another level of complexity. Low and Neutral Exploration profiles scrolled deeper than High Exploration, even though they interacted less overall. This pattern suggested that exploration can occur in multiple forms: active engagement with diverse elements or passive scanning through content. For interface evaluation, this distinction matters because designers often prioritize click-based or task-completion metrics, yet scrolling can represent a different form of engagement that deserves equal consideration.

These findings contribute to ongoing conversations in HCI and UX research about anticipating diverse user behaviors during design. Traditional methods, such as empirical usability testing, provide high-quality insight but are slow and costly to scale.[1,2] Post-launch tools like A/B testing and analytics offer systemic ways to assess design changes, but they require live users and traffic.[3,4] Our work demonstrates that LLMs can serve as personality-conditioned simulated users, providing a new pre-deployment tool that can augment established practices.

In addition, these results extend prior research showing that LLMs can simulate reasoning and decision making in non-interactive tasks[5–7] and role play as structured personas in multi-agent contexts with measurable actions, revealing that conditioning did not just influence which decisions were made but also the style of exploration itself. This underscores the potential of LLMs to model meaningful behavioral differences aligned with personality traits.

Our finding that personality conditioning shaped interaction patterns aligns with recent LLM personality research, though with important distinctions. Cohen *et al.*[16] found that Agreeableness and Extraversion significantly affect believability and goal achievement in LLM-simulated price negotiations through causal discovery methods, measuring how Big Five traits influence specific behavioral outcomes. They demonstrated that personality conditioning produces measurable behavioral variation in task-oriented contexts. However, their study focused on structured negotiation dialogues with clear endpoints, whereas our work examined exploratory UI navigation with less clearly-defined decision boundaries.

Limitations:

Several limitations constrain the generalizability of these findings. First, we lack human baseline data. While our results demonstrate that personality conditioning produces different behavioral patterns in LLMs, we cannot determine whether these patterns correspond to how humans with similar personality traits actually navigate the same interfaces. Lu *et al.*[17] found that LLMs achieved only 11.86% accuracy in multi-turn human behavior simulation in an e-commerce context. This emphasizes the gap between behavioral plausibility and accuracy.

Second, personality operationalization brings challenges. We condensed complex personalities into simplified 'exploration' profiles. Recent research by Xu *et al.*[18] demonstrated that persona conditioning introduces model-specific and context-dependent behavioral variations. They found that weakened agreeableness or conscientiousness significantly increased LLMs' susceptibility to unsafe outputs under bullying tactics. This suggests personality prompts may interact unpredictably with model architectures and situational contexts, activating behavioral patterns beyond simple trait expression. Our Neutral Exploration profile's hybrid strategy, which produced fewer actions than Low Exploration but covered more elements, may reflect similar complexity, where personality conditioning yields coherent but non-linear behavioral patterns whose psychological validity remains unclear.

Conclusion

Our study investigated whether LLMs could simulate personality-driven differences in user interaction when navigating web interfaces, and our findings confirmed that personality conditioning influenced both the frequency and style of engagement, supporting our central hypothesis. In particular, our study confirmed that High Exploration profiles interacted more broadly and frequently, Low Exploration profiles showed more limited engagement, while Neutral Exploration displayed a hybrid pattern, and both Neutral and Low Exploration highlighted the importance of passive engagement through scrolling.

These findings suggest that personality-conditioned LLM agents could become a valuable complement to interface evaluation processes, providing early insight into how different users

might behave through user testing, reducing the cost of user testing, and potentially reducing unethical use of human subjects. Several important questions remain, though. Do these patterns generalize to support more complex real-world applications? How closely do LLM-simulated behaviors align with those of human users who share similar personality traits as the LLMs? Which prompting strategies best sustain persona-consistent behavior across longer sessions?

We would like to propose that future research should combine simulations with human baselines, expand to richer task environments, and refine measurement methods in order to capture subjects' attention and engagement more directly. In addition, integrating personality-driven simulations with tools such as A/B testing and post-launch analytics could help determine the contexts where simulated users add the greatest value to the design and development process.

Acknowledgments

I would like to thank Dr. Siddharth Krishnan for his dedicated teaching in machine learning and artificial intelligence, and for keeping everything organized and on track throughout the project. His feedback was invaluable in shaping the direction of this research. I am also grateful to Dr. Plinio Zanini for his constructive feedback and technical support, which greatly strengthened the clarity and rigor of this paper.

References

1. Mao, J.-Y.; Vredenburg, K.; Smith, P. W.; Carey, T. The State of User-Centered Design Practice. *Commun ACM* **2005**, *48* (3), 105–109. https://doi.org/10.1145/1047671.1047677.
2. Karat, C.-M.; Campbell, R.; Fiegel, T. Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92*; ACM Press: Monterey, California, United States, **1992**; pp 397–404. https://doi.org/10.1145/142750.142873.
3. Quin, F.; Weyns, D.; Galster, M.; Silva, C. C. A/B Testing: A Systematic Literature Review. *J. Syst. Softw.* **2024**, *211*, 112011. https://doi.org/10.1016/j.jss.2024.112011.
4. Kohavi, R.; Deng, A.; Frasca, B.; Walker, T.; Xu, Y.; Pohlmann, N. Online Controlled Experiments at Large Scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*; ACM: Chicago, Illinois, USA, **2013**; pp 1168–1176. https://doi.org/10.1145/2487575.2488217.
5. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*; NIPS '22; Curran Associates Inc.: Red Hook, NY, USA, 2022; pp 24824–24837.
6. Dasgupta, I.; Lampinen, A. K.; Chan, S. C. Y.; Sheahan, H. R.; Creswell, A.; Kumaran, D.; McClelland, J. L.; Hill, F. Language Models Show Human-like Content Effects on Reasoning Tasks. arXiv July 17, **2024**. https://doi.org/10.48550/arXiv.2207.07051.
7. Rajani, N. F.; McCann, B.; Xiong, C.; Socher, R. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Korhonen, A., Traum, D., Màrquez, L., Eds.; Association for Computational Linguistics: Florence, Italy, **2019**; pp 4932–4942. https://doi.org/10.18653/v1/P19-1487.
8. Benharrak, K.; Zindulka, T.; Lehmann, F.; Heuer, H.; Buschek, D. Writer-Defined AI Personas for On-Demand Feedback Generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*; CHI '24; Association for Computing Machinery: New York, NY, USA, **2024**; pp 1–18. https://doi.org/10.1145/3613904.3642406.
9. Sparrenberg, L.; Schneider, T.; Deußer, T.; Koppenborg, M.; Sifa, R. Correcting Systematic Bias in LLM-Generated Dialogues Using Big Five Personality Traits. In *2024 IEEE International Conference on Big Data (BigData)*, **2024**; pp 3061–3069. https://doi.org/10.1109/BigData62323.2024.10825941.
10. Visconti, E.; Tsigkanos, C.; Nenzi, L. Automated Monitoring of Web User Interfaces. *ACM Trans Web* **2025**, *19* (2), 10:1-10:27. https://doi.org/10.1145/3708512.
11. Dhonde, S.; Joshi, A.; Jadhav, T.; Hote, S.; Jansari, Mr. M. Automated UI Testing on Frontend of a Web Application. *Int. J. Enhanc. Res. Sci. Technol. Eng.* **2024**, *13* (05), 221–227. https://doi.org/10.55948/IJERSTE.2024.0533.
12. Abb, L.; Rehse, J.-R. A Reference Data Model for Process-Related User Interaction Logs. arXiv July 25, **2022**. https://doi.org/10.48550/arXiv.2207.12054.
13. Mehri, S.; Yang, X.; Kim, T.; Tur, G.; Mehri, S.; Hakkani-Tür, D. Goal Alignment in LLM-Based User Simulators for Conversational AI. arXiv July 27, **2025**. https://doi.org/10.48550/arXiv.2507.20152.
14. Pegoraro, M.; Uysal, M. S.; Hülsmann, T.-H.; van der Aalst, W. M. P. Resolving Uncertain Case Identifiers in Interaction Logs: A User Study. arXiv November 21, **2022**. https://doi.org/10.48550/arXiv.2212.00009.
15. Gupta, N.; Comar, P. M. Quantifying Customer Interactions on ML Optimized Page Layouts. In P*roceedings of the International Conference on Advances in Social Networks Analysis and Mining*; ACM: Kusadasi, Turkiye, **2023**; pp 502–508. https://doi.org/10.1145/3625007.3627519.
16. Cohen, M. C.; Su, Z.; Kao, H.-T.; Nguyen, D.; Lynch, S.; Sap, M.; Volkova, S. Exploring Big Five Personality and AI Capability Effects in LLM-Simulated Negotiation Dialogues. arXiv August 20, **2025**. https://doi.org/10.48550/arXiv.2506.15928.
17. Lu, Y.; Huang, J.; Han, Y.; Yao, B.; Bei, S.; Gesi, J.; Xie, Y.; Zheshen; Wang; He, Q.; Wang, D. Prompting Is Not All You Need! Evaluating LLM Agent Simulation Methodologies with Real-World Online Customer Behavior Data. arXiv June 5, **2025**. https://doi.org/10.48550/arXiv.2503.20749.
18. Xu, Z.; Sanghi, U.; Kankanhalli, M. Bullying the Machine: How Personas Increase LLM Vulnerability. arXiv May 19, **2025**. https://doi.org/10.48550/arXiv.2505.12692.

Appendix A. Additional Data

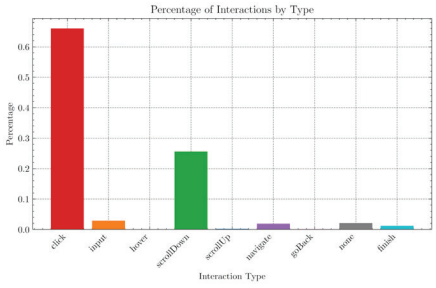

Figure A1: Normalized interaction type distribution across all profiles. Shows that LLMs tended to favor clicking and scrolling down regardless of their assigned profile.

DOI: 10.36838/IJHSR811.75

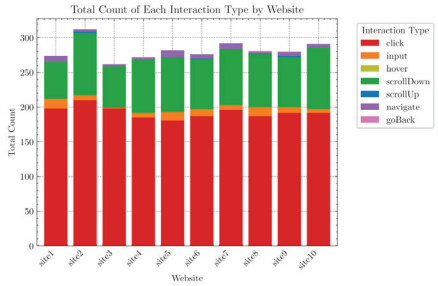

Figure A2: Interaction type counts by website. The type of website and the content shown caused variation in the way that LLMs acted and what actions were taken. This pattern replicates the way that humans interact differently depending on the website content.

Author

Isaac Robinson is a senior at Carlmont High School in California. He is a researcher focused on human-computer interaction and large language models. His work developing Mintable, a budgeting app for teens, motivated his investigation into how user interfaces are designed and how humans and AI systems interact with them.