OPeRA: A Dataset of Observation, Persona, Rationale, and Action for Evaluating LLMs on Human Online Shopping Behavior Simulation

Ziyi Wang[1], Yuxuan Lu[1], Wenbo Li[2], Amirali Amini[1], Bo Sun[1], Yakov Bart[1], Weimin Lyu[3], Jiri Gesi [4], Tian Wang[4] Jing Huang[4], Yu Su[5], Upol Ehsan[1], Malihe Alikhani[1], Toby Jia-Jun Li[6], Lydia Chilton[7], Dakuo Wang[1],

[1]Northeastern University, [2]University of Southern California, [3]Stony Brook University, [4]Independent Researcher, [5]Ohio State University, [6]University of Notre Dame, [7]Columbia University

Correspondence: wang.ziyi19@northeastern.edu, d.wang@northeastern.edu

Abstract

Can large language models (LLMs) accurately simulate the next web action of a specific user? While LLMs have shown promising capabilities in generating "believable" human behaviors, evaluating their ability to mimic real user behaviors remains an open challenge, largely due to the lack of high-quality, publicly available datasets that capture both the observable actions and the internal reasoning of an actual human user. To address this gap, we introduce OPeRA, a novel dataset of **O**bservation, **P**ersona, **R**ationale, and **A**ction collected from real human participants during online shopping sessions. OPeRA is the first public dataset that comprehensively captures: user personas, browser observations, fine-grained web actions, and self-reported just-in-time rationales. We developed both an online questionnaire and a custom browser plugin to gather this dataset with high fidelity. Using OPeRA, we establish the first benchmark to evaluate how well current LLMs can predict a specific user's next action and rationale with a given persona and <observation, action, rationale> history. This dataset lays the groundwork for future research into LLM agents that aim to act as personalized digital twins for human [1].

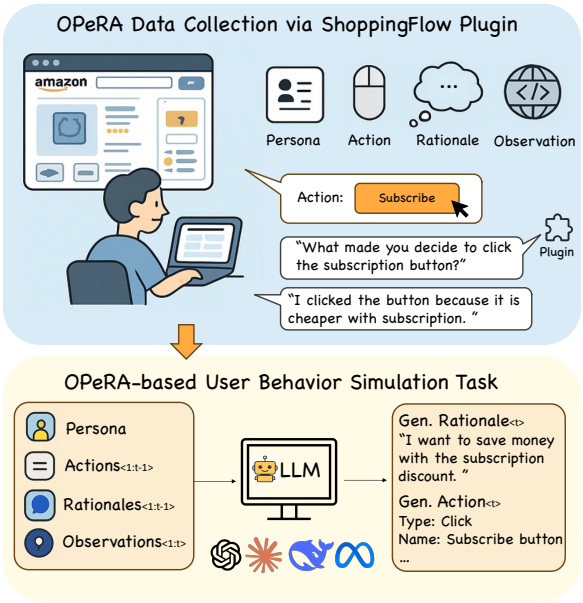

Figure 1: We developed ShoppingFlow plugin (Figure 2) to collect user shopping behavior over a four-week period, resulting in OPeRA-full dataset. This dataset comprises 692 sessions from 51 unique users, containing 28,904 real-user <action, observation> pairs and 604 user-annotated rationales (Figure 3). After post-processing, we obtained OPeRA-filtered, which includes 527 sessions, 5,856 <action, observation> pairs, and 207 rationales. We then benchmarked four LLMs' performance on user next action prediction task, results in Table 8).

1 Introduction

Large language model (LLM) agents have exhibited impressive performance across diverse tasks, including planning, reasoning, and acting in web-based environments (Xie et al., 2024; Yao et al., 2023; Jin et al., 2025). A promising frontier in this area is human behavior simulation, where LLM agents generate user-like action sequences on digital platforms (Chen et al., 2025a). These agents (i.e., role-playing agents) are increasingly used in applications such as UI/UX testing (Lu et al., 2025b), social science research (Park et al., 2024), accessibility testing (Taeb et al., 2024), and personal digital assistants (OpenAI, 2025b). Yet, while these agents can generate believable behavior, the more critical question remains: can they generate behavior that accurately aligns with real human?

Despite progress in agent-based behavior simulation (Park et al., 2023, 2024), current works still have limitations. First, most existing evaluations focus on aggregate outcomes (e.g., survey responses or end-task completions). These methods overlook the step-wise rationale and actions that underlie user behavior patterns (Chen et al., 2025a). For

[1]https://huggingface.co/datasets/NEU-HAI/OPeRA

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 43942–43960
July 2-7, 2026

example, Park et al. (2024) compared LLM agents' survey results with real humans by replicating various social science studies. Furthermore, many role-play agents rely solely on prompting without grounding in real human data training, which limits their accuracy and personalization. Although some recent efforts incorporate user behavior data via fine-tuning (Lu et al., 2025a), these datasets are often proprietary or lack critical detail, such as the reasoning or persona behind user actions.

Current open-source datasets for user behavior simulation fall short in several key aspects. First, most datasets record only sparse, decontextualized– or even synthetic–user actions. Some shopping datasets like Amazon-M2 or ECInstruct (Jin et al., 2024a,b) record only the isolated actions (e.g., purchases or clicks) with limited observation context. Others (Deng et al., 2024; Chen et al., 2024; Yao et al., 2022) use synthetic or third-party annotated behaviorial data, which lacks the individual behavior pattern and the authenticity. Additionally, few datasets provide step-level reasoning or persona information, despite prior work has shown that rationale can improve LLM agent's performance in behavior and decision modeling (Lu et al., 2025a; DeepSeek-AI et al., 2025). Similarly, user persona strongly correlates with behavioral patterns (Helmi et al., 2023), making persona essential for durable personalization experience.

To address these limitations, we introduce OPeRA, a dataset of **O**bservation, **Pe**rsona, **R**ationale, and **A**ction collected from real human users during online shopping. OPeRA provides rich, time-aligned logs of users' web browsing behavior, completed with self-reported rationales and detailed self-reported persona profiles. Unlike prior datasets, OPeRA captures not only **what** users do but also **why** they do it, enabling deeper insights into decision-making processes. This paper focus on the online shopping domain as a beginning point due to its everyday prevalence, complex decision flow, and strong ties to personalization (Mican and Sitar-Taut, 2020; Wei, 2016; Zhang et al., 2025b; Wang et al., 2025c; Zhang et al., 2025a). Online e-commerce environments like Amazon requires a user's multi-step interactions involving comparisons, trade-offs, and goal-directed behaviors in one shopping session, all of which are prime testbeds for studying the capacity of LLMs to simulate real human actions.

To collect the OPeRA dataset, we developed ShoppingFlow, a custom browser plugin that captures user interactions alongside corresponding web context and triggers rationale prompts at decision points, as shown in Figure 1. We also collect rich persona information through an online survey and an optional interview to include user profile information such as demographics, shopping styles, and personality traits.

The OPeRA-full contains 692 shopping sessions from 51 unique users, 28,904 <action, observation> pairs, and 604 human-annotated rationales. After post-processing, we also provide OPeRA-filtered, which includes 527 sessions, 5856 <observation, action> pairs and 207 rationales. OPeRA serves as the first benchmark dataset for evaluating LLM agents on **personalized** and **verifiable** user behavior simulation. We benchmark four state-of-the-art LLMs (GPT-4.1 (OpenAI, 2025a), DeepSeek-R1 (DeepSeek-AI et al., 2025), Claude-3.7 (Anthropic, 2025), and Llama-3.3 (Meta, 2024)) on OPeRA-test (a subset of the OPeRA-filtered) and analyze their ability to predict the next action and rationale of a specific user based on their persona and interaction history. These findings lay a foundation for future work on building LLM-powered digital twins capable of accurate and adaptive behavior modeling.

2 Related Works

2.1 LLMs for Human Behavior Simulation

Large Language Model agents can handle complex tasks (Wang et al., 2024b; Yao et al., 2023; Shinn et al., 2023; Wu et al., 2023; Zhang et al., 2024; Jia et al., 2024; Chen et al., 2025b; Wang et al., 2026), and researchers are using them as human proxies across domains, from social-science simulations (Park et al., 2023; Wang et al., 2024a; Sreedhar et al., 2025) and recommender-system evaluation (Wang et al., 2023) to UX testing (Lu et al., 2025b) and health-counsellor training (Louie et al., 2025). They even reproduce classic results in experimental psychology and economics, such as Milgram Shock Experiment (Aher et al., 2023).

In parallel, there have been works looking at generating synthesized personas based on text data, such as PersonaHub (Ge et al., 2024) and the approach by Shi et al. (2025), demonstrating promise in downstream modeling. Moreover, several approaches have integrated persona information to enrich behavioral simulation (Shao et al., 2023; Chuang et al., 2024; Shi et al., 2025). For example, Park et al. (2024) introduces an persona-

Dataset	Size	Task	O	Pe	R	A	Source
Amazon Review	571M	Review Prediction	✗	✗	✗	Purchase	User
ECInstruct-SA	10k	Sentiment Analysis	✗	✗	✗	Purchase	User
ECInstruct-REC	10k	Recommendation	✗	✗	✗	Purchase	User
Amazon-M2	3M	Recommendation	✗	✗	✗	Click	User
Repeat Buyers	54M	Buyer Prediction	✗	✓	✗	Click, Cart Favor, Purchase	User
Taobao	100M	Recommendation	✗	✗	✗	Click, Cart Favor, Purchase	User
YOOCHOOSE	9M	Purchase Prediction	✗	✗	✗	Click, Purchase	User
Shopping MMLU	3973	Recommendation	✗	✗	✗	Query, Click Purchase	User
Mind2Web	2350	Web Navigation	✓	✗	✗	Click, Hover, Type, Select	Annotator, GPT
GUI-WORLD	12k	GUI Understanding Instruction Follow	✓	✗	✗	Click, Paste, Search, Type	Annotator, Video
WebArena	812	Web Navigation Instuction Follow	✓	✗	✗	Click, Hover, Type, Tab Switching, Navigation	Annotator, GPT
WebShop	1600	Web Navigation	✓	✗	✗	Input, Click	Annotator
OPeRA-full OPeRA-filtered	692 527	All Above and User Behavior Simulation	✓	✓	✓	Basic Action, Semantic Action	User

Table 1: Properties of existing datasets compared to OPeRA. "O": Environment Observation. "Pe": User Persona. "R": Rationale behind action. "A": Action Space. "Source": Action Source.

grounded framework using qualitative interviews, enabling agents to accurately simulate individual preferences, attitudes, and behaviors, showing that incorporating personas improves the realism of simulated behavior.

Moreover, there is a growing trend for using LLM agents in online scenarios to produce human-like behaviors, such as Claude Computer Use (Anthropic, 2024), AutoGLM (Liu et al., 2024), Coco-Agent (Ma et al., 2024), Mobile-Agent-E (Wang et al., 2025b), and OpenAI Operator (OpenAI, 2025b), enabling interaction in more complex human-computer settings. Yet, existing research on LLM web agents predominantly focuses on optimizing and evaluating agents for task completion (Gur et al., 2023; Zhou et al., 2024a; He et al., 2024), training agents to learn the fewest, most direct steps–whereas real users behaviors normally contain richer and nonlinear paths. There remains a lack of work exploring accurate simulation of human behavior, particularly the modeling of personalized user behaviors. Motivated by this limitation, we propose OPeRA, which aims to better facilitate research on simulating realistic and personalized human behaviors in online scenarios.

2.2 User Online Behavior Datasets

Existing datasets containing user online behaviors can be broadly categorized into two sets, shown in Table 1. The first category consists of recommendation-oriented datasets which mostly records user interaction with product items (e.g., click, purchase), lacking of essential context. A commonly used example is the Amazon Review 2023 dataset (Hou et al., 2024), which contains over 571 million user review-writing behaviors and links them with product information. Follow-up studies, such as ECInstruct (Peng et al., 2024), have utilized the Amazon Review to work on tasks like user review sentiment analysis. Others, such as Amazon-M2 (Jin et al., 2024a), YOUCHOOSE (Ben-Shimon et al., 2015), Repeat Buyers (Liu et al., 2016), and Taobao (Zhu et al., 2018), capture richer shopping behaviors such as clicks, add-to-cart, and purchases. The Repeat Buyers dataset further offers basic user persona like gender and age. However, these datasets lack fine-grained user behaviors and contextual information needed to explain how actions and decisions are made at each stage of the user shopping journey, limiting their utility in user behavior simulation.

A separate line of datasets focuses on task com-

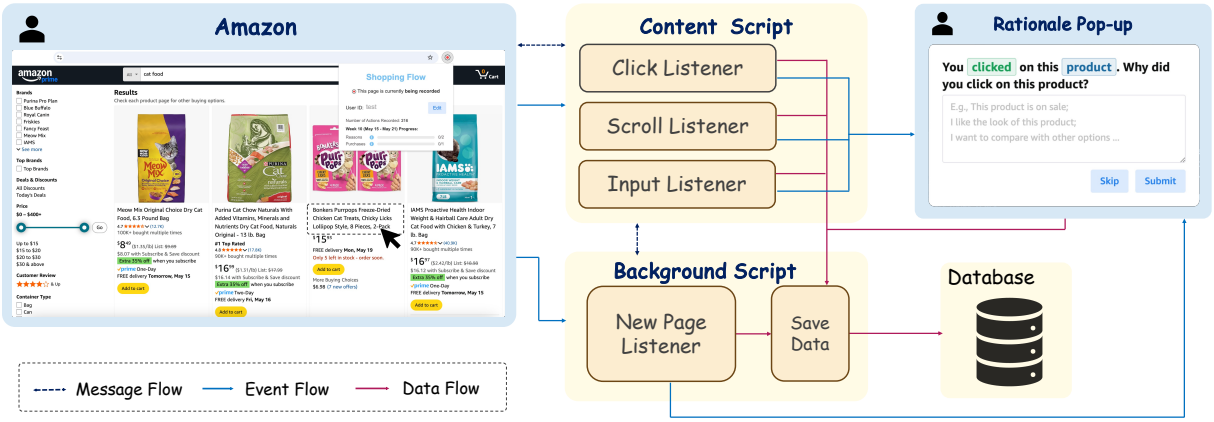

Figure 2: Pipeline of our Chrome Plugin, `ShoppingFlow`. `Content Script` detects click, scroll and input actions. `Background Script` detects page-related actions and handles data uploading. Rationale pop-up is triggered at a certain probability when certain action types are detected.

pletion, with environment observations for web agents tasks, such as `Mind2Web` (Deng et al., 2024), `GUI-WORLD` (Chen et al., 2024), `WebArena` (Zhou et al., 2024b), and `WebShop` (Yao et al., 2022). These datasets contain user interaction traces on website or mobile devices (e.g., clicks, typing) and HTML or screenshots. While useful for studies like instruction following, these datasets are built from annotator-generated or synthetic interaction logs. As a result, they lack both behavioral authenticity and personalization, limiting their realism.

There are other datasets and benchmarks which focus on evaluating LLM agents' reasoning ability using question-answering format, for example Mialon et al. (2023). While such tasks are valuable for evaluating models, these datasets are not suitable for simulating human behavior.

In summary, existing datasets capture certain aspects of user behaviors but lack the comprehensive data required for simulating real user behavior.

3 `OPeRA` Dataset

Our work presents a novel dataset `OPeRA` to advance NLP research in realistic user behavior simulation, particularly within the context of online shopping scenarios.

3.1 `OPeRA` Data Collection

3.1.1 Participant Recruitment

A total of 84 participants were recruited through snowball sampling (Goodman, 1961). During the pre-screening survey, candidate participants were required to self-identify if they were frequent Amazon customers and if they planned to make at least one purchase on Amazon in the next several weeks. All participants were required to install the `ShoppingFlow` Chrome plugin (details in Section 3.1.3), shop normally on the Amazon website for a four-week period, and participate in an online survey and an optional interview. Details of recruitment are in Appendix A.

3.1.2 Persona Information Collection

We collected detailed user persona information through a structured online survey and an optional semi-structured interview. All survey questions are designed based on established work (design details in Appendix B) to solicit consumer characteristics that have been shown to correlate with shopping behaviors. The survey consists of three main sections: **demographic information**, **shopping preferences**, and **personality traits**. Demographic information includes age, gender, education, occupation, income, residence, and self-description. Shopping preferences includes online shopping frequency, membership status, shopping habits, seasonality, advertising trust, review engagement, delivery influence, and an adapted eight-item Consumer Styles Inventory (CSI) (Nayeem and Marie-IpSooching, 2022). Personality traits are measured using the Big-Five Inventory (Goldberg, 1992) as well as a self-reported MBTI personality (Myers et al., 1998).

Following prior work on persona information collection (Park et al., 2024), we invited participants to attend a 20-minute optional semi-structured interview about their **demographic information, personal background, and online shopping preferences**, which aimed to better contextualize the personalized decision-making processes behind users' online shopping actions. The

interview design description and the protocol used are provided in Appendix C.

3.1.3 Shopping Behavior Collection

To support the construction of the dataset, we designed `ShoppingFlow`, a Chrome extension that automatically captures user behaviors and contextual web observations during Amazon shopping sessions (shown in Figure 2).

The plugin includes two main scripts: a `Content Script` that runs within the Amazon page to log user interactions (including inputs, clicks, and scrolls) with timestamps, target elements, and HTML using custom parsing rules (Appendix D); and a `Background Script` that tracks page-level events like reloads and navigation. All data was securely uploaded to Amazon S3. To capture rationale, the plugin randomly triggers pop-ups (8% chance) asking users to explain their actions (question design in Appendix E).

3.1.4 Post-Processing

To ensure data quality and to **protect user privacy,** we applied a multi-step post-processing procedure. We configured the plugin to not record any personally identifiable information (PII), such as the user login page, account profile page, or the checkout details. In addition, we designed a rule-based automated detection and pattern matching script to mask any PII unavoidably contained in a page (e.g., username in navigation bar), including usernames, zip codes, addresses, specific workplaces, and payment details, before any human touch. Lastly, we manually checked the data to ensure there is no PII in the dataset.

The actual purchase information is not collected since it is in the checkout detail page. Instead, we infer a "purchase" action via click actions on "proceed to checkout," "buy now," and "set subscription" buttons. We associate the inferred purchase action with the corresponding product information during post-processing.

In addition, the raw user data is a stream of continuous user action sequences that do not separate different shopping sessions. Thus, our team segmented user actions using a two-step strategy based on temporal intervals and purchase signals. Sessions were first split using a time threshold and then further segmented at purchase intention events (i.e., clicking on "proceed to checkout / buy now / set subscription / add to cart" button). Detailed explanation about the selection of the threshold can be found in Appendix F. Finally, sessions with fewer than five actions were discarded to remove trivial or non-informative behaviors as a prior work reported meaningful sessions typically contain at least six to seven interactions (Wang et al., 2025a).

To reduce noise and improve the quality of behavioral data, some actions that occurred on uncommon or rarely visited pages or do not reflect meaningful intent are removed. Moreover, we filtered out clicks on non-interactive areas such as the background, and further filtered actions involving Amazon Rufus.

To support behavior modeling and evaluation under a more tractable setting, we follow prior work (Lu et al., 2025a) and similarly define a simplified action $\mathcal{A}$ space consisting of key interaction types. Specifically, we retain three high-level actions that are both semantically meaningful and commonly observed: input, click, and terminate. Within the "click" category, we further differentiate between several subtypes, shown in Table 3. This abstraction reduces complexity while preserving the structure of user behaviors, enabling more stable behavior simulation.

3.2 Dataset Details

An overview of the dataset is presented in Figure 3. Of all the users, 51 contributed at least one shopping session. In total, the `OPeRA-full` contains 692 sessions, 28,904 <action, observation> pairs, and 604 rationale annotations. The `OPeRA-filtered` contains 527 sessions, 5,856 <action, observation> pairs, and 207 rationales. Table 2 presents the statistics. Table 3 shows the click type distribution.

User Persona For participant i, the persona is represented by P_i, comprising two components: a structured survey and interview. Both focusing on their demographics, personality or personal background, and shopping preferences.

User Action Traces Each shopping session j includes users' web interactions on the shopping website. The action trace is represented by $A_j = \{a_1, ..., a_T\}$, where $a_t \in \mathcal{A}$ represents the user's action at step t in the action space $\mathcal{A}$. The Action space $\mathcal{A}$ includes `click`, `scroll`, `input`, `navigate`, `tab activate`. Each action is assigned a unique identifier (UUID) and is timestamped to preserve the exact temporal order.

In particular, for each click action, the corresponding CSS selector of the element is

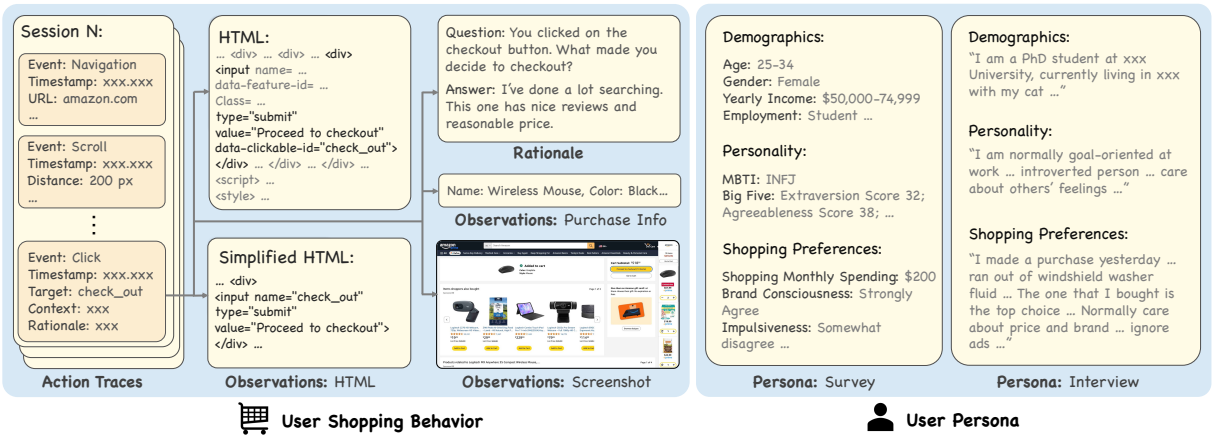

Figure 3: OPeRA Dataset Overview. The dataset comprises four major components: **action traces**, **web observations**, **rationales**, and **user personas**. Each shopping session is a sequence of timestamped actions. Each action is paired with a corresponding web observation, which includes: the full HTML of the interacted webpage, a simplified HTML with key elements, a screenshot, and product information for purchases (if applicable). The rationale is a natural language explanation of why the user performed the action. The persona contains detailed user profiles collected from surveys and interviews, covering demographics, personality traits, and shopping preferences.

provided to uniquely identify the clicked target. Additionally, semantic identifiers are assigned to click actions to indicate their functional context, such as interactions with products in search results (e.g., `semantic_id`: "`search_result.product_name`") or interactions with other commonly used page elements. This semantic id enables downstream models to have a clearer recognition and observation of user behaviors when utilizing the dataset. Similarly, scrolling actions are presented with their start and end positions, supporting analyses of how users explore page content.

Rationale Alongside the user actions, rationales are provided for some actions, $R_j = \{r_1, \ldots, r_T$, where r_k is a nullable string describes the user's rationale for specific action a_k, explicitly capturing the underlying motivations or thought processes of users. These insights enable the downstream models to have a thorough understanding of why users make particular choices, offering a deeper view of user behavior and the reasoning process.

Web Observation In addition to action traces and rationales, the observation of the web context at each action step in session j is captured, represented by $O_j = \{o_0, ..., o_T\}$. Each o_t includes the **HTML** content and a **screenshot**[2]. The HTML content also contains annotations indicating the viewability of each clickable element, as well as relevant page metadata (such as product name, product price, etc.) at time t. To reduce noise and storage while keeping the HTML structure, we provides the a simplified version of the HTML content containing key elements for frequently encountered pages, such as search results, product detail pages, and shopping carts.

Additionally, if a purchase is made during a session, the final purchase information, including the price, product title, product options, and Amazon Standard Identification Number (ASIN), are recorded to facilitate potential downstream tasks such as recommendation.

Action Type	# of Full	# Filtered
Scroll	19,217 (66.5%)	–
Click	5,253 (18.1%)	5,051 (86.3%)
Tab Activate	1,945 (6.7%)	–
Navigate	1,901 (6.6%)	–
Text Input	606 (2.1%)	597 (10.2%)
Terminate	–	208(3.6%)
Total	28,904	5,856

Table 2: Action type distribution.

4 Tasks and Experiments

4.1 Tasks

We show how OPeRA can be leveraged to evaluate LLM's ability to simulate consumer behavior in online shopping.

[2]The experiments presented in this paper did not involve the use of screenshot data.

Click Type	Count	Percentage
review	1052	20.8%
search	763	15.1%
product_option	700	13.9%
product_link	537	10.6%
other	449	8.9%
purchase	321	6.4%
nav_bar	283	5.6%
page_related	198	3.9%
quantity	191	3.8%
suggested_term	182	3.6%
cart_side_bar	145	2.9%
cart_page_select	139	2.8%
filter	91	1.8%

Table 3: Click type distribution in OPeRA-filtered. Detailed descriptions in Appendix G

Metric	Value
# of Session	527
Avg. # of Action	11.11
Avg. # of Input	1.13
Avg. # of Click	9.58
Avg. # of Terminate	0.39

Table 4: OPeRA-filtered dataset statistics per-session

Next Action Prediction The next action prediction task aims to model the user's next action based on previous behaviors following the definition of previous work (Deng et al., 2024; Lu et al., 2025a). Given a history of actions $\{a_1, \cdots, a_{t-1}\}$ in shopping session j, corresponding web contexts $\{o_1, ..., o_t\}$, rationale $\{r_1, ..., r_t\}$[3], and the consumer profile P_i, the model is tasked to predict the immediate next action a_t, learning a function of the form:

$$a_t = F_{action}(a_{1...t-1}, r_{1...t-1}, o_{1...t}, P_i)$$

4.2 Experiments

4.2.1 Experiment Setup

From the OPeRA-filtered, we furthur construct the test set **OPeRA-test** by randomly sampling 15 out of 51 users and randomly sampling 90 sessions from these users, resulting in 992 actions. We evaluate four state-of-the-art LLMs, including two open-source models, Llama-3.3-70B-Instruct and DeepSeek-R1, and two proprietary models, GPT-4.1 and Claude-3.7-Sonnet.

All models are evaluated in a zero-shot, prompt-based setting without fine-tuning. Prompt templates are provided in Appendix I.

To investigate how different input factors affect model behavior, we conducted a series of ablation studies. Specifically, we excluded persona information (`w/o persona`) and additionally removed the history rationale from the input (`w/o rationale`).

4.2.2 Evaluation

To assess the accuracy of generated user actions, we apply an exact match criterion: a prediction is correct only when all required components align with the ground truth. Specifically, for click actions, the clicked target name must match. For input actions, this includes identifying the input field and generating exact input text.

In addition to exact match accuracy, we evaluate the model's ability to classify action types. We report weighted F1 for high-level action types (`click`, `input`, `terminate`). Given the highly imbalanced nature of user behavior distributions, we also report macro F1 to highlight performance across all classes regardless of frequency.

To further examine fine-grained prediction capabilities, we evaluate the weighted F1 score for `click` subtypes. This captures whether the model not only predicts that a user will click but also understands the specific type of click (e.g., `review`, `product_link`, `purchase`).

Finally, given the goal-oriented nature of online shopping, we assess the model's ability to predict session outcomes. Each session ends in either a `click on purchase-related button` or a `terminate` action. We evaluate the model's accuracy and F1 score on these terminal actions to understand whether it can correctly capture users' long-term goals and decision-making processes.

4.2.3 Results and Analysis

4.3 Main Results

The results for next action prediction are presented in Table 8. Among all models, GPT-4.1 achieves the strongest overall performance. It obtains the highest action generation accuracy of 22.06%, a macro F1 score of 48.78% on action type prediction, and a weighted F1 score of 44.47% on click type prediction. For session outcome prediction, GPT-4.1 also shows solid performance with

[3]Note: Rationale annotations are sparse, whereas actions and observations are fully recorded at each time step.

Model	Action Gen. (Accuracy)	Action Type (Macro F1)	Click Type (Weighted F1)	Session Outcome (Weighted F1)
GPT-4.1	21.51	**48.78**	**44.47**	47.54
w/o persona	**22.06**	45.55	43.45	**58.47**
w/o rationale	21.28	34.93	42.63	51.17
DeepSeek-R1	14.75	27.37	35.12	46.36
w/o persona	15.52	27.43	33.86	48.86
w/o rationale	15.74	27.16	32.65	47.92
Claude-3.7	10.75	31.58	27.27	43.52
w/o persona	10.75	25.33	22.76	43.10
w/o rationale	10.08	26.06	20.29	43.10
Llama-3.3	8.31	24.29	19.99	36.64
w/o persona	8.31	23.69	18.59	33.21
w/o rationale	8.76	23.60	19.22	34.19

Table 5: Evaluation of actions in next action prediction task. We report four metrics here to assess model performance (Full results can be found in Appendix H): Action Generation Accuracy measures the exact-match accuracy of the predicted next action; Action Type Macro F1 evaluates the model's ability to predict the correct high-level action category (e.g., input, click, terminate); Click Type Weighted F1 captures the performance of predicting the specific type of click actions; Session Outcome Weighted F1 reflects how well the model can predict the final outcome of a session, where each session ends either in a purchase or a terminate action. "Claude-3.7": Claude-3.7-Sonnet, "Llama-3.3": Llama-3.3-70B-Instruct. All metrics are reported as percentages (%). Instance size $n = 902$.

58.47% F1. This strong performance may be attributed to its large context window, which facilitates better processing of complex interaction histories. DeepSeek-R1 performs moderately well across tasks. It achieves a maximum action generation accuracy of 15.74% and obtains solid outcome prediction, possibly due to its strong reasoning abilities. However, the performance may be limited by the model's 128k context length. Claude-3.7 shows modest performance. Its action generation accuracy is around 10.75%, and action type prediction is generally weaker than GPT-4.1 and DeepSeek. Nonetheless, it achieves relatively good outcome prediction, suggesting some robustness in capturing high-level user intent. LLaMA-3.3 underperforms across all metrics, with action generation accuracy of only 8.76%. Its lower performance may be due to the smaller model size (70B) and shorter context length (128k).

The role of persona information varies across models. While adding persona information does not consistently improve exact action generation accuracy, it generally enhances the model's performance on action type and click type prediction across all model families. This suggests that persona information provides useful priors about user preferences and behavior patterns, helping the model better classify action semantics. However, its limited effect on action generation accuracy implies that simply including persona in the prompt may introduce noise rather than help. Current models have limited ability to deeply integrate persona into step-level decision-making. This highlights potential room for improvement in personalized user modeling.

Additionally, removing historical rationales consistently leads to performance degradation across most models and metrics, particularly in outcome prediction. This confirms that rationale information serves as valuable intermediate supervision, guiding the model to align its decisions with plausible user intent. We also note several outliers in the results. For instance, LLaMA-3.3 does not consistently benefit from rationale inputs. This may be due to its smaller model size and limited capacity to leverage additional contextual signals effectively.

4.4 Error Analysis

As shown in Table 6, the majority of model failures are attributed to incorrect button click predictions. In addition, models frequently struggle to accurately generate input or termination actions. Even in cases where the model successfully identifies an input action, it often fails to reproduce the correct

Error Type	GPT	R1	Claude	LLaMA
Didn't Terminate	35 (3.9%)	39 (4.3%)	40 (4.4%)	40 (4.4%)
Didn't Click	49 (5.4%)	21 (2.3%)	33 (3.7%)	27 (3.0%)
Didn't Input	50 (5.5%)	70 (7.8%)	55 (6.1%)	74 (8.2%)
Input Wrong Field	0 (0.0%)	0 (0.0%)	1 (0.1%)	0 (0.0%)
Input Wrong Text	26 (2.9%)	6 (0.7%)	19 (2.1%)	2 (0.2%)
Click Wrong Button	548 (60.8%)	633 (70.2%)	657 (72.8%)	684 (75.8%)

Table 6: Error type breakdown across models with count and percentage.

Action Type	Ground-Truth	GPT	R1	Claude	LLaMA
Click	786 (87.14%)	819 (90.80%)	865 (95.90%)	843 (93.46%)	862 (95.57%)
Input	76 (8.43%)	54 (5.99%)	21 (2.33%)	48 (5.32%)	3 (0.33%)
Terminate	40 (4.43%)	29 (3.22%)	5 (0.55%)	0 (0.00%)	0 (0.00%)
Other	0 (0.00%)	0 (0.00%)	11 (1.22%)	11 (1.22%)	37 (4.10%)

Table 7: Distribution of predicted and ground truth action types with count and percentage.

search query.

Table 7 further highlights the discrepancy between the predicted and ground-truth distributions of action types. Notably, the "terminate" action, despite its presence in the ground-truth data, is rarely predicted by most models (except for GPT-4.1). This mismatch suggests a potential bias in current LLMs to be optimized for completing the shopping task (i.e., the purchase), rather than simulating realistic user behavior, which often includes early session termination.

5 Conclusion

This paper introduces OPeRA, a comprehensive online shopping behavior dataset specifically designed to advance the development and evaluation of LLM-based agents for simulating user behavior. By capturing full shopping trajectories, including action traces, web observations, user personas, and explicit rationales, the dataset provides a verifiable, personalized resource for user behavior modeling. We define a suite of evaluation tasks and conduct a comprehensive analysis across four state-of-the-art LLMs. Our results highlight both the promise and current limitations of LLM agents in simulating realistic user behavior. While certain models demonstrate plausible rationale and action prediction under simple setups, there remains substantial room for improvement, especially in handling complex decision flows and deeper personalization.

Limitations

This study evaluates LLM agents under a simplified setup, following prior work (Lu et al., 2025a), adopting a reduced action space and a coarse-grained session segmentation strategy. In particular, actions such as scrolling and page navigation are omitted. This simplification serves to manage the complexity of the task: while text-based LLMs are robust in processing structured language data, they might struggle to accurately model continuous UI-based actions like scrolling. We hope that future models capable of richer, multimodal reasoning may better handle these complexities.

Second, although we collect screenshots alongside HTML data for every user interaction, the experiments in this paper do not incorporate visual signals. This is primarily because existing LLM agents are not yet robustly equipped to interpret raw visual UI elements from screenshots in conjunction with structured web content. We envision that future work can leverage this visual information to interpret user decisions.

Looking forward, the OPeRA dataset enables a wide range of future research directions beyond those explored in this paper (e.g. personalized recommendation). Furthermore, simulation based on OPeRA offers a scalable framework for generating synthetic interaction data, which could be valuable for applications like website design and adaptive user interaction. We hope this dataset serves as a foundation for more realistic, personalized, and interpretable user behavior modeling.

References

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Anthropic. 2024. Claude 3.5 sonnet technical overview. https://www.anthropic.com/news/claude-3-5-sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Anthropic. 2025. Claude 3.7 sonnet and claude code. Accessed: 2025-04-06.

Živilė Baubonienė and Gintarė Gulevičiūtė. 2015. E-commerce factors influencing consumers 'online shopping decision. *Social technologies*, 5(1):62–73.

David Ben-Shimon, Alexander Tsikinovsky, Michael Friedmann, Bracha Shapira, Lior Rokach, and Johannes Hoerle. 2015. Recsys challenge 2015 and the yoochoose dataset. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 357–358.

Chaoran Chen, Bingsheng Yao, Ruishi Zou, Wenyue Hua, Weimin Lyu, Yanfang Ye, Toby Jia-Jun Li, and Dakuo Wang. 2025a. Towards a design guideline for rpa evaluation: A survey of large language model-based role-playing agents. *arXiv preprint arXiv:2502.13012*.

Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, et al. 2024. Gui-world: A dataset for gui-oriented multimodal llm-based agents. *arXiv preprint arXiv:2406.10819*.

Yiwei Chen, Soumyadeep Pal, Yimeng Zhang, Qing Qu, and Sijia Liu. 2025b. Unlearning isn't invisible: Detecting unlearning traces in llms from model outputs. *arXiv preprint arXiv:2506.14003*.

Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V Frigo, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2024. Beyond demographics: aligning role-playing llm-based agents using human belief networks. *arXiv preprint arXiv:2406.17232*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Preprint*, arXiv:2501.12948.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2024. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics*, pages 148–170.

Izzeddin Gur, Hiroki Furuta, Austin V. Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2023. A Real-World WebAgent with Planning, Long Context Understanding, and Program Synthesis. In *The Twelfth International Conference on Learning Representations*.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. WebVoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6864–6890, Bangkok, Thailand. Association for Computational Linguistics.

Arief Helmi, Rita Komaladewi, Vita Sarasi, and Ledy Yolanda. 2023. Characterizing young consumer online shopping style: Indonesian evidence. *Sustainability*, 15(5):3988.

Jianwei Hou and Kevin Elliott. 2021. Mobile shopping intensity: Consumer demographics and motivations. *Journal of Retailing and Consumer Services*, 63:102741.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024. Soul: Unlocking the power of second-order optimization for llm unlearning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4276–4292.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, Haoyu Han, Hanqing Lu, Zhengyang Wang, Ruirui Li, et al. 2024a. Amazon-m2: A multilingual multi-locale shopping session dataset for recommendation and text generation. *Advances in Neural Information Processing Systems*, 36.

Yilun Jin, Zheng Li, Chenwei Zhang, Tianyu Cao, Yifan Gao, Pratik Jayarao, Mao Li, Xin Liu, Ritesh Sarkhel, Xianfeng Tang, et al. 2024b. Shopping mmlu: A massive multi-task online shopping benchmark for large language models. *arXiv preprint arXiv:2410.20745*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Guimei Liu, Tam T Nguyen, Gang Zhao, Wei Zha, Jianbo Yang, Jianneng Cao, Min Wu, Peilin Zhao, and Wei Chen. 2016. Repeat buyer prediction for e-commerce. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 155–164.

Xiao Liu, Bo Qin, Dongzhu Liang, Guang Dong, Hanyu Lai, Hanchen Zhang, Hanlin Zhao, Iat Long Iong, Jiadai Sun, Jiaqi Wang, et al. 2024. Autoglm: Autonomous foundation agents for guis. *arXiv preprint arXiv:2411.00820*.

Ryan Louie, Ifdita Hasan Orney, Juan Pablo Pacheco, Raj Sanjay Shah, Emma Brunskill, and Diyi Yang. 2025. Can llm-simulated practice and feedback upskill human counselors? a randomized study with 90+ novice counselors. *Preprint*, arXiv:2505.02428.

Yuxuan Lu, Jing Huang, Yan Han, Bingsheng Yao, Sisong Bei, Jiri Gesi, Yaochen Xie, Zheshen, Wang, Qi He, and Dakuo Wang. 2025a. Prompting is not all you need! evaluating llm agent simulation methodologies with real-world online customer behavior data. *Preprint*, arXiv:2503.20749.

Yuxuan Lu, Bingsheng Yao, Hansu Gu, Jing Huang, Jessie Wang, Laurence Li, Jiri Gesi, Qi He, Toby Jia-Jun Li, and Dakuo Wang. 2025b. Uxagent: An llm agent-based usability testing framework for web design. *Preprint*, arXiv:2502.12561.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2024. Coco-agent: A comprehensive cognitive mllm agent for smartphone gui automation. *arXiv preprint arXiv:2402.11941*.

Meta. 2024. Llama-3.3-70b-instruct.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. *Preprint*, arXiv:2311.12983.

Daniel Mican and Dan-Andrei Sitar-Taut. 2020. Analysis of the factors impacting the online shopping decision-making process. *Studia Universitatis Babes-Bolyai*, 65(1):54–66.

Isabel Briggs Myers, Mary H. McCaulley, Naomi L. Quenk, and Allen L. Hammer. 1998. *MBTI Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator*, 3rd edition. Consulting Psychologists Press, Palo Alto, CA.

Tahmid Nayeem and Jean Marie-IpSooching. 2022. Revisiting sproles and kendall's consumer styles inventory (csi) in the 21st century: A case of australian consumers decision-making styles in the context of high and low-involvement purchases. *Marketing*, 7(2):7–17.

Nurul Zarirah Nizam and Jaafar Abdullah Jaafar. 2018. Interactive online advertising: The effectiveness of marketing strategy towards customers purchase decision. *International Journal of Human and Technology Interaction (IJHaTI)*, 2(2):9–16.

OpenAI. 2025a. Gpt-4.1.

OpenAI. 2025b. Introducing operator.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. *Preprint*, arXiv:2411.10109.

Bo Peng, Xinyi Ling, Ziru Chen, Huan Sun, and Xia Ning. 2024. ecellm: Generalizing large language models for e-commerce from large-scale, high-quality instruction data. *arXiv preprint arXiv:2402.08831*.

Gyan Prakash, Pankaj Kumar Singh, and Rambalak Yadav. 2018. Application of consumer style inventory (csi) to predict young indian consumer's intention to purchase organic food products. *Food quality and preference*, 68:90–97.

Jing-bo Shao, Zhen-zhen Li, and Ming-ye Hu. 2014. The impact of online reviews on consumers' purchase decisions in online shopping. In *2014 International Conference on Management Science & Engineering 21th Annual Conference Proceedings*, pages 287–293. IEEE.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Yimin Shi, Yang Fei, Shiqi Zhang, Haixun Wang, and Xiaokui Xiao. 2025. You are what you bought: Generating customer personas for e-commerce applications. *arXiv preprint arXiv:2504.17304*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Mastercard SpendingPulse. 2023. Holiday spending 2023. Accessed: March 5, 2025.

George B Sprotles and Elizabeth L Kendall. 1986. A methodology for profiling consumers' decision-making styles. *Journal of Consumer Affairs*, 20(2):267–279.

Karthik Sreedhar, Alice Cai, Jenny Ma, Jeffrey V Nickerson, and Lydia B Chilton. 2025. Simulating cooperative prosocial behavior with multi-agent llms: Evidence and mechanisms for ai agents to inform policy decisions. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 1272–1286, New York, NY, USA. Association for Computing Machinery.

Maryam Taeb, Amanda Swearngin, Eldon Schoop, Ruijia Cheng, Yue Jiang, and Jeffrey Nichols. 2024. AXNav: Replaying Accessibility Tests from Natural Language. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–16, New York, NY, USA. Association for Computing Machinery.

Dakuo Wang, Ting-Yao Hsu, Yuxuan Lu, Limeng Cui, Yaochen Xie, William Headean, Bingsheng Yao, Akash Veeragouni, Jiapeng Liu, Sreyashi Nag, et al. 2025a. Agenta/b: Automated and scalable web a/btesting with interactive llm agents. *arXiv preprint arXiv:2504.09723*.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2024a. User behavior simulation with large language model based agents. *Preprint*, arXiv:2306.02552.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024b. Sotopia-π: Interactive learning of socially intelligent language agents. *arXiv preprint arXiv:2403.08715*.

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. 2023. Recmind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296*.

Zhenhailong Wang, Haiyang Xu, Junyang Wang, Xi Zhang, Ming Yan, Ji Zhang, Fei Huang, and Heng Ji. 2025b. Mobile-agent-e: Self-evolving mobile assistant for complex tasks. *arXiv preprint arXiv:2501.11733*.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Jing Huang, Jiri Gesi, Xianfeng Tang, Chen Luo, Yisi Sang, Hanqing Lu, Manling Li, et al. 2026. Trajectory2task: Training robust tool-calling agents with synthesized yet verifiable data for complex user intents. *arXiv preprint arXiv:2601.20144*.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Jing Huang, and Dakuo Wang. 2025c. Customer-r1: Personalized simulation of human behaviors via rl-based llm agent in online shopping. *arXiv preprint arXiv:2510.07230*.

Liying Wei. 2016. Decision-making behaviours toward online shopping. *International Journal of Marketing Studies*, 8(3):111–121.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. Travelplanner: A benchmark for real-world planning with language agents. *arXiv preprint arXiv:2402.01622*.

Shunyu Yao, Howard Chen, John Yang, and Karthik R. Narasimhan. 2022. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D Lee, Wotao Yin, Mingyi Hong, et al. 2024. Revisiting zeroth-order optimization for memory-efficient llm fine-tuning: A benchmark. *arXiv preprint arXiv:2402.11592*.

Yimeng Zhang, Jiri Gesi, Ran Xue, Tian Wang, Ziyi Wang, Yuxuan Lu, Sinong Zhan, Huimin Zeng, Qingjun Cui, Yufan Guo, et al. 2025a. See, think, act: Online shopper behavior simulation with vlm agents. *arXiv preprint arXiv:2510.19245*.

Yimeng Zhang, Tian Wang, Jiri Gesi, Ziyi Wang, Yuxuan Lu, Jiacheng Lin, Sinong Zhan, Vianne Gao, Ruochen Jiao, Junze Liu, et al. 2025b. Shop-r1: Rewarding llms to simulate human behavior in online shopping via reinforcement learning. *arXiv preprint arXiv:2507.17842*.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024a. WebArena: A Realistic Web Environment for Building Autonomous Agents. *Preprint*, arXiv:2307.13854.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024b. Webarena: A realistic web environment for building autonomous agents. *Preprint*, arXiv:2307.13854.

Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. 2018. Learning tree-based deep model for recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1079–1088.

A Data Collection Implementation Details

We recruited 84 participants via snowball sampling. The participants are pre-screened to ensure that they meet the following criteria: at least 18+ years old, based in the U.S., English speaker and have used (or plan to use) Amazon website to make a purchase in the past (future) couple of weeks.

The participant incentive structure consists with: a) $5 for completing the online survey (10 mins) and b) $10 for participating in an optional 20-mins interview to discuss about their personal background. c) $5 per one qualified week for participants who have complete one or more purchase sessions, with at least two rationale recorded by the `ShoppingFlow` plugin. In addition, we provide $10 as a bonus to participants who can successfully complete four or more qualified weeks of data collection in a row. All incentives were delivered as Amazon digital gift cards to their emails.

During the data collection process, no personally identifiable information was retained. To ensure data privacy, we don't collect data on sensitive pages like checkout or account page. Meanwhile, we implemented a script to anonymize sensitive details in recorded context data, such as zip codes, names, and addresses. Any screenshots that unintentionally contained identifiable information were reviewed and removed prior to dataset release. Furthermore, a research assistant conducted a post-screening review to confirm the exclusion of identifiable information.

This study was conducted in compliance with the Institutional Review Board (IRB) guidelines at Northeastern University.

B Survey Design

The survey consists of three main sections: **demographic information**, **shopping preferences**, and **personality traits**.

Demographic information significantly influences consumer behavior (Hou and Elliott, 2021). The survey collects age, gender, education level, occupation, family income, location of residence, and a two-sentence self-description.

Shopping preferences section asks for participants' online shopping frequency and whether they have a paid membership or not. In addition, we include 12 questions inspired by previous literature, all with a 5-point Likert scale answer (Likert, 1932). These questions include four shopping habits items, seasonality (SpendingPulse, 2023), tendency of believing in advertisements (Nizam and Jaafar, 2018), habits of reading product reviews (Shao et al., 2014), and the influence of delivery (Baubonienė and Gulevičiūtė, 2015) and an 8-items consumer styles inventory (CSI) adapted to the online shopping context (Sprotles and Kendall, 1986; Helmi et al., 2023; Prakash et al., 2018; Nayeem and Marie-IpSooching, 2022), including: brand loyalty, price conciousness, perfectionism and high-quality conciousness, impulsiveness, confusion by overchoice, brand consciousness, recreational consciousness, and novelty and fashion conciousness.

Personality traits section utilizes the Big-Five Personality Inventory with five core dimensions: Openness to experience, conscience, extrovertism, agreeableness, and neuroticism (Goldberg, 1992). A self-reported MBTI is also included (Myers et al., 1998).

B.1 Survey Questions

B.1.1 Demographic Information

Q1: Gender.

[Male; Female; non-binary]

Q2: Age.

[Under 18; 18-24; 25-34; 35-44; 45-54; 55-64; 65+]

Q3: Which city do you live?

Q4: What is your highest level of education ?

[High school diploma or lower; Bachelors' degree or current college student; Graduate degree or current grad student (MA, MS, MBA, etc.); Doctoral degree or current doctoral student (PhD, JD, MD, DDS etc.); Prefer not to say]

Q5: Do you live alone or live together with others, if so who are they? (Optional)

Q6: What was your household before-tax income during the past 12 months? If you are a student, what's your allowance or stipend?

[Less than $25,000; $25,000-$49,999; $50,000-$74,999; $75,000-$99,999; $100,000-$149,999; $150,000 or more; Prefer not to say]

Q7: What best describes your employment status over the last three months?

Q8: What best describes your employment status over the last three months?

[Full-time employee; Part-time employee; Self-employed; Unemployed and looking for work; Student; Retired; Other:

Q9: Use two sentences to describe yourself. Example 1:"I am a machine learning researcher

at a startup focusing on autonomous driving. My daily work include developing and optimizing deep learning models for perception, sensor fusion, and decision-making in self-driving vehicles." Example 2: "I am a PhD student in computer science, specializing in AI for healthcare. I frequently conduct experiments, analyze medical datasets, and collaborate with doctors to ensure our models are clinically interpretable. I also attend academic seminars, present my findings at conferences, and participate in lab meetings to refine research directions. "

B.1.2 Shopping Preferences

Q10: How often do you shop online?

[More than three times a week; Once to twice a week; Once every couple of weeks; Less than once a month]

Q11: How much money (in US dollars) do you spend on online shopping per month? (Not including food or delivery services)

Q12: Do you have a paid Amazon Prime membership?

[Yes; No]

From Q13 to Q24, all items use the same response scale: [Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree]

Q13: I tend to shop more during holidays (e.g. Black Friday, holiday sales).

Q14: Online ads attract my attention and are a good source of information.

Q15: I usually do a lot of research (e.g. reading online reviews) before making a purchase.

Q16: I prioritize delivery speed and delivery fee of the product.

Q17: Getting high-quality online products is very important for me.

Q18: The more expensive online product brands are usually my choice.

Q19: The more I learn about online products, the harder it seems to choose the best.

Q20: I shop quickly for online products, buying the first product or brand I find that seems good enough.

Q21: Once I find a brand I like, I stick with it.

Q22: I would buy a new or different brand of product just to see what it is like.

Q23: I enjoy shopping for online products just for the fun of it.

Q24: I look carefully to find the best value for money when shopping online.

B.1.3 Personality Traits

Big Five Test: read the statements carefully and indicate to what extent you agree of disagree. From Q25 to Q74, all items use the same response scale: [Very Inaccurate, Moderately Inaccurate, Neither Accurate Nor Inaccurate, Moderately Accurate, Very Accurate]

Q25: Am the life of the party.

Q26: Feel little concern for others.

Q27: Am always prepared.

Q28: Get stressed out easily.

Q29: Have a rich vocabulary.

Q30: Don't talk a lot.

Q31: Am interested in people.

Q32: Leave my belongings around.

Q33: Am relaxed most of the time.

Q34: Have difficulty understanding abstract ideas.

Q35: Feel comfortable around people.

Q36: Insult people.

Q37: Pay attention to details.

Q38: Worry about things.

Q39: Have a vivid imagination.

Q40: Keep in the background.

Q41: Sympathize with others' feelings.

Q42: Make a mess of things.

Q43: Seldom feel blue.

Q44: Am not interested in abstract ideas.

Q45: Start conversations.

Q46: Am not interested in other people's problems.

Q47: Get chores done right away.

Q48: Am easily disturbed.

Q49: Have excellent ideas.

Q50: Have little to say.

Q51: Have a soft heart.

Q52: Often forget to put things back in their proper place.

Q53: Get upset easily.

Q54: Do not have a good imagination.

Q55: Talk to a lot of different people at parties.

Q56: Am not really interested in others.

Q57: Like order.

Q58: Change my mood a lot.

Q59: Am quick to understand things.

Q60: Don't like to draw attention to myself.

Q61: Take time out for others.

Q62: Shirk my duties.

Q63: Have frequent mood swings.

Q64: Use difficult words.

Q65: Don't mind being the center of attention.
Q66: Feel others' emotions.
Q67: Follow a schedule.
Q68: Get irritated easily.
Q69: Spend time reflecting on things.
Q70: Am quiet around strangers.
Q71: Make people feel at ease.
Q72: Am exacting in my work.
Q73: Often feel blue.
Q74: Am full of ideas.
Q75: What is your MBTI personality type? (Optional)

C Interview Design

The interview includes question sections of **demographic info, detailed personal background, and online shopping habits and preferences**. The interviews encourages participants to elaborate on their personas through open responses, thus facilitating a more nuanced understanding of their individual experiences and decision-making processes. For example, in the personal narrative section, participants were encouraged to describe a typical day and share how they perceive themselves. Similarly, in the online shopping-related section, they were asked to describe a recent purchase session on Amazon, providing concrete examples of their shopping habits for downstream models' understanding and simulation of consumer behavior. This includes pre-purchase research activities, in-session shopping behaviors, engagement with review content, and attitude on advertisement.

C.1 Interview Protocal

Demographics
Introduction: Can you tell me a bit about yourself? What kind of work do you do? Where do you live? Do you live alone or with family?
Personal Background
Daily Life: You mentioned your work/study. What does a typical day look like for you?
Work Activities: Can you tell me more about [job/study]? What are your main responsibilities or daily tasks?
After-Work Activities: What do you usually do after work?
Self-Perception: How would you describe yourself?
Online Shopping Preferences
Recent Purchase: Can you recall a recent Amazon purchase? What was the reason for shopping? What were you looking for? How did you find and decide on the product?
Pre-Shopping Activities: Did you research before making the purchase?
During-shopping Shopping: How long did it take to decide? How many products did you compare before choosing?
Decision Factors: What mattered most in your choice? Do your priorities change based on category (e.g., style for clothing, brand for electronics, price for essentials)?
Reviews: Did you read any reviews before purchasing? What information were you looking for? If not, why?
Advertisements: Do you notice sponsored products? How do ads influence your decisions?

D Web Parser Design

This section introduces the parser designed to process and simplify web pages, enabling downstream LLMs to better understand and interact with the page content. The parser is guided by a framework called a recipe, which consists of a set of JSON-based rules tailored to specific web pages. These recipes are created through a combination of manual rule-writing and automated assistance from GPT.

Each recipe uses CSS selectors to identify and extract key HTML elements that are important for user interaction (e.g., search boxes, filter options) or containing important semantic information (e.g., product prices, availability status).

We designed recipes for several common page types, such as search result pages, product detail pages, and checkout pages. When parsing a page, the parser extracts these key elements and annotates them with a unique semantic ID. For instance, the ID `refinements.colors.red` denotes an element that filters results to red-colored items when clicked. These semantic IDs help the model understand both the structure and the function of the elements. Clickable elements are further annotated with visual markers to inform the downstream model of their interactivity.

To simplify the HTML, the parser removes all irrelevant content and styling, retaining only the extracted key elements. This results in a clean, minimal HTML structure that is easier for the model to analyze and reason over.

E Rationale Prompt Design

The plugin employs context-aware pop-up questions to collect user rationale across different interaction types. The question design follows a hierarchical structure based on event types and specific interaction targets.

Click Events

- *Click on subscription setup button*: You clicked on the set up now button. Can you tell us why you subscribed to this product?
- *Click on buy now button*: You clicked on the "buy now" button. Why did you do that?
- *Click on add to cart button*: You clicked on the "add to cart" button. Why did you decide to add this product to your cart?
- *Click on search button*: You clicked on the "search" button. Why did you make this search?
- *Click on filters*: You clicked on this filter. Why did you use this filter?
- *Click on product options*: You clicked on this product option. Why did you click this product option?
- *Click on checkout button*: You clicked on the "checkout" button. What made you decide to checkout?
- *Click on decrease quantity button*: You clicked on the "decrease quantity" button. Why did you click this button?
- *Click on increase quantity button*: You clicked on the "increase quantity" button. Why did you click this button?
- *Click on product list*: You clicked on this product. Why did you click on this product?
- *Click on other area*: We noticed that you just had a click action. Why did you do that?

Scroll Events

- *Page scrolling*: We saw that you scrolled up/-down this page. What are you looking for?

Navigation Events

- *Back / forward navigation*: Why did you decide to return to this page?

Tab Switch Events

- *Tab activation*: Why did you leave and come back to this tab?

F Segmentation Interval Threshold Justification

To determine the time interval threshold for session segmentation, we analyzed the distribution of inter-action time gaps from the first week of user activity. From observation, approximately 98% of intervals were shorter than 4 minutes, and 99% were shorter than 78 minutes. Choosing the 99th percentile (78 minutes) as the threshold balances session granularity, avoiding both excessive fragmentation and excessively long sessions that mix unrelated behaviors.

In addition, behavioral analysis showed that users typically interact with the site for 50 to 150 actions before reaching a natural session boundary, such as a purchase or exit, which means around every 50 to 150 interactions, there is probably a session termination or purchase signal. This observation supports the chosen 99% threshold as both statistically and behaviorally grounded.

G Click Type Description

This section provides detailed descriptions for the click type categories used in the OPeRA dataset, as presented in Table 3. The following categories were established through empirical analysis of user click action distributions.

review (20.8%): Clicks on review-related elements, including review images, star ratings, review filters, etc.

search (15.1%): Clicks on search button or search box.

product_option (13.9%): Clicks on product options such as size selectors, color options etc.

product_link (10.6%): Clicks on product images, product titles, or product links that navigate users to product detail pages.

other (8.9%): Miscellaneous clicks that do not fall into the above predefined categories.

purchase (6.4%): Clicks on purchase-intention elements including "Add to Cart", "Buy Now", "Subscribe", and "Checkout".

nav_bar (5.6%): Clicks on navigation bar elements such as category menus, amazon logo etc.

page_related (3.9%): Clicks on pagination controls, carousel navigation buttons that control page content display.
quantity (3.8%): Clicks on quantity adjustment controls including increase/decrease buttons and item deletion buttons.
suggested_term (3.6%): Clicks on search suggestions.
cart_side_bar (2.9%): Clicks on elements in shopping cart sidebar.
cart_page_select (2.8%): Clicks on selection elements such as item checkbox in the cart page.
filter (1.8%): Clicks on filtering elements including price filters, brand filters, rating filters, and other product refinement controls.

H Next Action Prediction Experiment Results

Table 8 shows the full results of next action prediction task.

I Experiment Prompt Design

Below are the two prompts for action prediction task and joint rationale and action generation task:

```
PROMPT_FOR_ACTION_PREDICTION = """
<IMPORTANT>
Your task is to predict the immediate next
    ↪ action of a shopper.
You need to pretend that you are a real user
    ↪ shopping on amazon.com.
The history action, rationale, context and the
    ↪ user persona will be provided to you.
Ensure your prediction follows natural behavior
    ↪ sequences (e.g., users may click a
    ↪ search box before typing, type a query
    ↪ before clicking search)
</IMPORTANT>

# Action Space

An action is represented in JSON format, and
    ↪ there are four primary types of actions:

#### 1. `input`:
Type text into an input field.
{
    "type": "input",
    "name": "input_name",
    "text": "input_text"
}

#### 2. `click`:
Click on a button or clickable element
    ↪ identified by `name`.
{
    "type": "click",
    "name": "clickable_name",
}

#### 3. `terminate`:
When you are unsatisfied with the current
    ↪ search result and you don't want to buy
    ↪ anything, use `terminate` to indicate
    ↪ that you want to close the browser
    ↪ window and terminate the task.
{
    "type": "terminate"
}

# Rationale
Rationale is the reason why the user takes the
    ↪ action. Some of the rationale is
    ↪ provided to you.

# Context
Your context will be the HTML of the amazon
    ↪ page you are looking at. Some
    ↪ interactable elements will be added a
    ↪ unique "name" attribute, which you can
    ↪ use to identify the element to interact
    ↪ with (click or input).

# Persona
The user persona reflects the user's
    ↪ demographics, personality, and shopping
    ↪ preference. First identify which aspects
    ↪ of the persona might be relevant to the
    ↪ current shopping context, then consider
    ↪ them only if they naturally align with
    ↪ the ongoing shopping journey. DO NOT
    ↪ RELY ON IT.

# Output Format
You need to predict the next action. Your
    ↪ output should follow a strict JSON
    ↪ format:
{
    "type": "<type>",
    ...
}

<IMPORTANT>
OUTPUT A SINGLE JSON OBJECT, NOTHING ELSE.
</IMPORTANT>
"""


"""
```

Model	Action Gen. (Accuracy)	Action Type (Weighted F1)	Action Type (Macro F1)	Click Type (Weighted F1)	Outcome (Accuracy)	Outcome (Weighted F1)
GPT-4.1	21.51	**85.04**	**48.78**	**44.47**	38.89	47.54
w/o persona	**22.06**	82.32	45.55	43.45	55.55	**58.47**
w/o rationale	21.28	83.13	34.93	42.63	53.33	51.17
DeepSeek-R1	14.75	81.99	27.37	35.12	51.11	46.36
w/o persona	15.52	81.72	27.43	33.86	**56.67**	48.86
w/o rationale	15.74	81.66	27.16	32.65	53.33	47.92
Claude-3.7	10.75	83.41	31.58	27.27	52.22	43.52
w/o persona	10.75	82.28	25.33	22.76	50.00	43.10
w/o rationale	10.08	81.08	26.06	20.29	47.78	43.10
Llama-3.3	8.31	80.69	24.29	19.99	34.44	36.64
w/o persona	8.31	78.59	23.69	18.59	28.89	33.21
w/o rationale	8.76	80.23	23.60	19.22	31.11	34.19

Table 8: Evaluation of actions in next action prediction task. "Claude-3.7": Claude-3.7-Sonnet, "Llama-3.3": Llama-3.3-70B-Instruct. All metrics are reported as percentages (%). Instance size $n = 902$.