Can LLM Agents Simulate Multi-Turn Human Behavior? Evidence from Real Online Customer Behavior Data

Anonymous ACL submission

Abstract

Recent research shows that LLM Agents can generate "believable" human behaviors via prompt-only methods, and such agents have been increasingly adopted in downstream applications. However, existing evaluation of these agents only focuses on qualitative believability (whether human raters think they are accurate), leaving open questions of whether LLM agents can accurately generate step-by-step actions mimicking a particular human's behavior in a multi-turn interaction task. In this work, we take shopping as a case study and present the first large-scale quantitative evaluation of state-of-the-art LLMs' ability to accurately simulate human behavior. Using real-world data from 31,865 online shopping sessions containing 230,965 user actions, our evaluation reveals that prompt-based LLMs (DeepSeek-R1, Llama, Claude) achieve only 11.86% accuracy in generating human actions, highlighting **a substantial gap in actual behavioral accuracy**. Through experiments, we also showcase that strategies as simple as fine-tuning LLMs on real human click-through data augmented with synthesized reasoning traces can greatly enhance models' performance. The fine-tuned Qwen2.5-7B achieves 17.26% action generation accuracy and 33.86% F1 score on final purchase prediction, representing substantial improvements of 5.4% and 13.85% over prompt-only baselines. This work establishes the first rigorous benchmark for human behavior simulation and provides actionable insights for developing more accurate LLM agents for future downstream applications.

1 Introduction

Recent advances in large language models (LLMs) have enabled the simulation of human behavior across a range of applications, including web automation (Gur et al., 2023; Zhou et al., 2024), social interaction behaviors (Park et al., 2023), interpersonal trust behaviors (Xie et al., 2024), and user interface interactions (Taeb et al., 2024). These developments have sparked a growing interest in developing **LLM Agent**, agentic AI systems that uses LLMs to simulate human behaviors (Chen et al., 2025a), in many domains, such as social science simulation (Park et al., 2024), UI/UX testing (Lu et al., 2025a), and expert annotation simulation (Chen et al., 2025b). Specifically, in the online shopping domain, researchers have used LLM Agents as virtual customers to test website features (Lu et al., 2025a), conduct automated A/B testing on design variants (**?**), and evaluate agentic AI systems (Sun et al., 2025). These applications build on the finding that LLM Agents are able to simulate "believable" human behavior (Park et al., 2023, 2024), whether those behaviors are aligned with step-by-step human actions remains unclear. Thus, a fundamental question remains unanswered: **How accurate can LLMs truly replicate human behavior?**

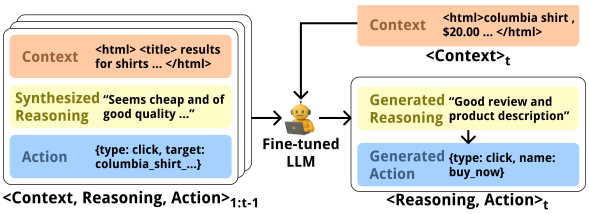

Figure 1: Overview of the next action prediction task. The model takes the currently observed $\langle$**context**$\rangle_t$ and a sequence of previous $\langle$**context, reasoning, action**$\rangle_{1:t-1}$ as input, and generates the next $\langle$**reasoning, action**$\rangle_t$ as output. Because the real-world human behavior dataset does not have groundtruth reasoning, we generate **synthesized reasoning trace** to complement the $\langle$context, action$\rangle$ pair.

In this paper, we focus specifically on the human behavior simulation task that is to **generate the next action the user is most likely to perform in a multi-turn interaction session, based on the current observation and the history of past actions**. For instance, in an online shopping scenario, the model observes the current webpage context (e.g., a product list) and the user's action history (e.g., previous clicks or queries), and generates the next plausible action a human would take (e.g., add some product to the shopping cart).

Existing evaluations (Park et al., 2023) of human behavior simulation primarily emphasize subjective measures of "**believability**" ("how much people feel it is like a human") rather than the objective "**accuracy**"

("how much it acts like a human"). The most relevant works that measure the objective model accuracy focus only on the final outcome of a task (e.g., purchasing the final product or not (Yao et al., 2023), or ultimately trusting the partner or not (Xie et al., 2024)), without examining whether the intermediate decision and action sequences align with those of actual humans. Consequently, the field currently lacks a robust and quantitative understanding for assessing LLMs at the process-centric, action-level simulation of human behaviors.

To bridge this gap, we take online shopping as a case study and provide the first systematic evaluation of SOTA LLMs' accuracy in process-centric, action-level behavior simulation tasks. We leverage a large-scale, real-world dataset consisting of 31,865 user click-through sessions from 3,526 users on an online shopping platform[1]. Each shopping session (Figure 1) comprises a series of timestamp-aligned ⟨context, action⟩pairs, where the context reflects the webpage observed by the user (e.g., product views, filter states), and the action denotes user inputs such as clicks, searches, or session termination actions. In total, the dataset has 230,965 user actions, and the final outcomes of the sessions include 4,432 purchase actions and 27,433 session termination actions. This dataset enables us to rigorously evaluate how accurately various LLMs can generate human-like behaviors at the action level. We evaluate different models on the next action generation task, benchmarking both the accuracy of generated actions throughout the session and the F1 score for final session outcome prediction (i.e. purchase or not), following protocols similar to existing work on modeling online shopping behavior.

Beyond evaluation, our dataset uniquely positions us to **fine-tune** LLMs to enhance their accuracy in behavior simulation tasks. While prior work has primarily relied on prompt-based approaches, we show that simply fine-tuning the model with user click-through data leads to significantly better accuracy in action generation and session outcome prediction. Furthermore, drawing inspiration from reasoning-augmented modeling (DeepSeek-AI et al., 2025), we hypothesize that exposing models to intermediate reasoning traces–even if synthetically generated–can enhance their ability to simulate human behavior. To test this hypothesis, we augment our dataset with synthesized reasonings from action traces using Claude 3.5 Sonnet and fine-tune models using this augmented data (⟨`context, action, reasoning`⟩ triplets) to learn how to generate not only accurate actions but also the underlying reasoning. Our results show that this reasoning-augmented fine-tuning further boosts model performance, highlighting the importance of modeling not just what humans do, but also why they do it. Taken together, our results provide the first quantitative evidence that out-of-the-box LLMs cannot accurately simulate action-level human behavior in realistic settings, and our ablation and error analyses provide insights to support future research.

In summary, this paper has three main contributions:

- We conduct the **first quantitative and process-centric evaluation** of LLMs' ability to simulate human web action behaviors using real-world online shopping data.
- We show that out-of-the-box LLMs cannot accurately predict human behavior in the next action prediction setting, while simple fine-tuning achieves substantially better performance.
- We demonstrate that **fine-tuning LLMs with synthesized reasoning traces further enhances their accuracy**, highlighting the importance of modeling not only what humans do but also why they do it for faithful human behavior simulation.

2 Related Works

2.1 Simulation of Human Behavior with LLM

The core function of the emerging LLM agent systems is their capability of generating human behaviors, in which a model takes a static user persona (e.g., preferences, demographics, or shopping habits), and the session data (e.g., a sequence of actions) as input to generate the next user action. Such systems have been extensively utilized and tested to simulate human behavior in a variety of scenarios. Park et al. (2023) simulated social behavior using generative agents in a virtual town, producing "believable" interactions. Xie et al. (2024) studied LLM agents in Trust Games to assess their ability to model human trust behavior. Park et al. (2024) used LLMs to simulate responses from 1,052 individuals in a social science survey. To simulate UI interaction, Lu et al. (2025a) proposed UXAgent, enabling LLMs to operate within web environments for simulated usability testing. Collectively, these studies underscore the growing potential of LLM-driven simulations to model and simulate complex, interpretable human behaviors.

However, the evaluation of these works remains limited in scope. Some focus on the subjective believability of process-centric action traces. For instance, Lu et al. (2025a) conducted qualitative interviews to assess participants' perceptions of the realism of their UXAgent system. Similarly, Park et al. (2023) proposed an evaluation framework that identified emergent social behaviors among generative agents. On the other hand, works that pursue objective evaluation often do so in a single-shot, outcome-centric manner. Zhou et al. (2024) introduced WebArena, a controlled environment for benchmarking web agents based on task completion rates. ReAct (Yao et al., 2023) measured success rates in simulation environments, overlooking the accuracy of model mimicing step-by-step human behavior. To date, no prior work has focused on **objectively evaluating model-generated step-by-step actions**–that is, assessing whether a model's sequence of decisions faithfully aligns with human behavior at each step.

[1]Platform name omitted for anonymity

2.2 Reasoning in Human Behavior Simulation

Building on the chain-of-thought prompting strategy (Wei et al., 2023), numerous studies have incorporated reasoning mechanisms into human behavior simulation. Park et al. (2023) pioneered agents equipped with reflection modules that synthesize memory and social context to support introspective decision-making. ReAct (Yao et al., 2023) prompted models to generate reasoning traces and actions separately, improving task success rates in online shopping and gaming environments. Gur et al. (2023) proposed WebAgent, which uses a dedicated reasoning model to plan sub-steps in web browsing tasks, enhancing control and planning in real-world browser simulations. Beyond single-agent reasoning, systems such as ChatDev (Qian et al., 2024) and RepoAgent (Luo et al., 2024) adopt multi-agent setups, where agents with specialized roles (e.g., programmers, testers) engage in collaborative dialogues via structured prompts. These communicative exchanges support more robust collective reasoning, demonstrating how coordination between agents can improve the quality of generated reasoning traces.

However, the aforementioned works incorporate reasoning using prompt-only approaches for the action generation task. Whether reasoning can improve performance in *fine-tuning* settings remains an open question. Although datasets for reasoning behind model actions exist (e.g., for conversational agents (Dongre et al., 2025)), there is currently no ready-to-use dataset specifically designed for human behavior simulation that includes both human actions and the corresponding reasoning behind those actions. To address a similar cold-start problem in reinforcement learning, DeepSeek-AI et al. (2025) constructed a small-scale dataset of long reasoning traces by synthesizing examples through few-shot prompting, followed by reflection and verification. Inspired by their methodology, we apply a similar strategy to synthesize reasoning traces in the online shopping domain. This enables us to investigate whether integrating reasoning into fine-tuning can enhance a model's ability to accurately simulate human behavior.

3 Method

3.1 Task Definition

In this section, we formally define the proposed human behavior simulation task: in the online shopping scenario, a shopping session is represented as a sequence of user actions $a_{1\ldots t\ldots N}$, always starting with a `search action` and concluding with either a `product purchase action` or a `termination action` (i.e., the user closing the browser window).

At each time step t, the model is tasked with generating both **the reasoning** r_t and **the action** a_t. The model input includes the current context c_t (what the user currently observes), a sequence of previous contexts (what the user has observed) $c_{1\ldots t-1}$, a sequence of previous actions $a_{1\ldots t-1}$ (what the user has done), and the corresponding synthetic reasoning trace $r_{1\ldots t-1}$ (why the user had that action) within the same session. Formally, the model learns a function f such that:

$$f(c_{1\ldots t}, a_{1\ldots t-1}, r_{1\ldots t-1}) = r_t, a_t$$

Observation Context The **Context** (or "observation space") of the web agent encompasses all available information on a webpage, including textual content, metadata, visual elements, and structural data. This context is designed to reflect how a human perceives and interprets a webpage, allowing the agent to process relevant features and perform tasks such as navigation, information retrieval, and interaction with page elements.

Previous research has explored various approaches to defining the context. Some methods rely on manually structured information parsing (Yao et al., 2022), while others utilize raw HTML representations (Gur et al., 2023) or accessibility tree embeddings (Zhou et al., 2024). However, both approaches have notable limitations: manually structured information parsing requires significant human effort to develop parsing rules (Yao et al., 2022), while raw HTML representations often contain extraneous information (e.g., CSS) that is not directly observable to human users, which may bias the LLM performance.

To ensure the adaptability to unseen websites, we define and implement a **simplified HTML** format as our context representation. This format removes non-relevant elements such as scripts, CSS, and purely visual components while preserving essential structural information. Using simplified HTML offers several advantages compared to custom defined formats (domain specific language, DSLs) or markdown-based representations: (1) important structural elements, such as lists and tables, remain intact, and (2) LLMs are already familiar with the HTML format, eliminating the need to redefine common elements like "button" and "input".

The LLM agent needs to refer to specific elements within the HTML, such as identifying the exact button it intends to click. Since there is no built-in method to uniquely identify HTML elements, prior work has proposed approaches like assigning sequential IDs to elements (Koh et al., 2024) or manually defining descriptive names for elements, such as `searchbox` (Yao et al., 2022). Following prior work (Lu et al., 2025a), we assign a unique hierarchical *name* in natural language to each interactable element, including links, buttons, and input fields. This name is constructed by incorporating the names of all parent nodes. For instance, if a `<a>` tag named `view_product` resides within a `<div>` named `columbia_shirt`, the resulting hierarchical name will be `columbia_shirt.view_product`.

Reasoning Reasoning trace refers to a natural language description that articulates the reasoning and explanation behind an action. For example, if the context is a search results page displaying a list of clothes, and the generated action is clicking on the `"4 stars and up"` product review filter, the generated reasoning might be: *"I want to find a comfortable piece*

of clothing, so I'm looking for options with high ratings." In our domain, the reasoning trace is missing from any real-world online shopping data, so we use LLM to synthesize it (Section 3.2). The generated reasoning provides insight into the model's thinking process, enhancing the transparency of the model's generations.

Action Previous research has explored various approaches to defining action spaces, including task-specific semantic actions such as "searching", "adding items to a cart", and "making purchases" (Yao et al., 2022), as well as browser-level interactions like `typing` and `clicking` (Lu et al., 2025a).

To ensure the adaptability of our framework beyond online shopping tasks, we define the action space at the level of **raw browser actions**, rather than at the level of task-specific semantics. The action space of our model consists of three fundamental browser operations: `click`, `type_and_submit`, and `terminate`. This abstraction allows the system to generalize across different environments while maintaining task flexibility.

3.2 Synthesized Reasoning Trace

Reasoning traces are crucial for understanding users' action choices but are difficult to collect; thus, they are often not available in behavioral datasets. We employ a reasoning synthesis pipeline to generate them using an LLM. To guide the reasoning generation process, we provide the LLM with the observation context and the corresponding action. Additionally, we record real human customers' think-aloud shopping sessions (Eccles and Arsal, 2017) as in-context learning examples. We then prompt the LLM to generate a free-text reasoning explaining the user's decision. Following recent works on reasoning-augmented LLMs (Wei et al., 2023; DeepSeek-AI et al., 2025), **the synthesized reasoning is not intended to replicate the actual human thought process**. Instead, its purpose is to enhance the model's predictive accuracy by providing structured intermediate representations that help the LLM better align actions with contextual cues. This approach ensures that the reasoning traces are coherent with the observed actions while improving the model's behavioral fidelity and explainability.

3.3 Model Architecture

To incorporate these enriched action traces, we build on existing pre-trained LLMs as our base models. The input to the model consists of two components: **(1) a sequence of historical contexts** (what the user observed), and **(2) the corresponding actions and generated reasoning trace** (what the user did and why). The model is trained to generate the next reasoning corresponding action, enabling it to simulate human behavior with greater fidelity.

During the **training stage**, the model receives the full sequence of a user session—including context, synthetic reasoning, and action—as a single concatenated input. The training objective is to minimize the next-token prediction loss for the reasoning and action tokens, while the loss for the context tokens is masked out. Subsequently, in the **evaluation stage**, the model works in an autoregressive manner: it is provided with historical context and past reasoning traces and actions, and is asked to first generate the reasoning for the next action; then, based on the generated reasoning, it generates the next action.

4 Experiments

4.1 Dataset Construction

Our training dataset was constructed using data from one of the largest e-commerce platforms globally [2]. The dataset contains 31,865 user sessions from 3,526 users in the online shopping scenario, comprising 230,965 user actions. The session's final outcome includes 4,432 purchase actions and 27,433 session termination actions. We leveraged our data synthesis pipeline detailed in Section 3.2 to generate the synthetic reasoning for each action based on the context using Claude-3.5-Sonnet. **The dataset was derived from traffic logs of a small group of users who explicitly opted into a beta testing feature and consented to data collection during the process. Users who opted out of data collection were excluded from the dataset.** The data was anonymized to remove any personally identifiable information during processing. Due to the sensitive nature of the data and to protect user privacy, we are unable to release the dataset.

We extracted pairs of user actions and contextual information from the cleaned data. These raw logs were then structured into the standardized format defined in Section 3.1.

4.2 Evaluation and Metrics

Evaluation Dataset We used a subset of the dataset that was not used during training as the test set, ensuring that no user sessions in the test set were seen by the model during fine-tuning. To create test cases, we took the second and all subsequent actions within each session, excluding the first action, since it lacks any preceding context. For each test case, the model was provided with the historical context, along with all previous actions and corresponding synthetic reasoning traces in the same session. The model is **first tasked with predicting the next reasoning, and then, based on the reasoning it generated**, it produces the corresponding action.

Prompt-Only Baseline To assess pre-trained models' ability in predicting human behavior, we evaluated a set of instruction-tuned LLMs under the in-context learning (ICL) setting (a.k.a., prompt-based setting). Specifically, we used several variants of Claude, LLaMA, and Mistral as representatives for general-purpose pre-trained LLMs, along with DeepSeek-R1 as a representative for

[2] Platform name omitted for anonymity

Model	Generated Next Action			Session Outcome		
	Accuracy	%Δ vs Base	v.s. DS-R1	F1 Score	%Δ vs Base	v.s. DS-R1
Open-Source Models						
DeepSeek-R1	11.86%	-	-	20.01%	-	-
Llama 3.1 8B	5.05%	-	-6.81%	10.87%	-	-9.14%
Llama 3.1 70B	8.19%	-	-3.67%	12.69%	-	-7.32%
Mixtral 8x7B	5.41%	-	-6.45%	13.16%	-	-6.85%
Qwen2.5-7B	4.25%	-	-7.61%	11.94%	-	-8.07%
Qwen2.5-3B	3.91%	-	-7.95%	10.87%	-	-9.14%
Qwen2.5-1.5B	3.27%	-	-8.59%	7.94%	-	-12.07%
Mistral-7B-v0.3	4.25%	-	-7.61%	11.27%	-	-8.74%
Llama 3.2 3B	2.93%	-	-8.93%	8.60%	-	-11.41%
Llama 3.2 1B	3.71%	-	-8.15%	3.09%	-	-16.92%
Proprietary Models						
Claude 3.5 Haiku	9.18%	-	-2.68%	14.77%	-	-5.24%
Claude 3 Opus	6.78%	-	-5.08%	15.08%	-	-4.93%
Claude 3 Sonnet	8.42%	-	-3.44%	17.40%	-	-2.61%
Claude 3.5 Sonnet	9.72%	-	-2.14%	15.91%	-	-4.10%
Claude 3.5 Sonnet v2	11.69%	-	-0.17%	18.54%	-	-1.47%
Claude 3.7 Sonnet	9.34%	-	-2.52%	12.81%	-	-7.20%
Fine-tuned Models						
Qwen2.5-7B	16.67%	-	4.81%	26.92%	-	6.91%
+ reasoning	**17.26%**	3.54%	5.40%	33.86%	25.78%	13.85%
Mistral-7B-v0.3	14.17%	-	2.31%	17.99%	-	-2.02%
+ reasoning	15.84%	11.79%	3.98%	30.12%	67.43%	10.11%
Llama-3.2-3B	9.31%	-	-2.55%	4.73%	-	-15.28%
+ reasoning	15.77%	69.39%	3.91%	**33.99%**	618.60%	13.98%

Table 1: Model performance. The table shows model accuracy in two tasks: the process-centric **action generation** task and the **outcome**-centric final purchase prediction task of the session. More models' performances are in the Appendix. DS-R1: performance comparision with DeepSeek-R1.

reasoning LLMs. To preserve user privacy, we run open-source model on our own GPU cluster and

In this setup, each model was provided with the historical context and previous user actions from the session, and prompted to generate both the next reasoning and the next action. The generated actions were used to compute macro accuracy for evaluation. These baselines reflect the commonly adopted approach of using powerful LLMs without domain-specific fine-tuning, allowing us to directly assess the impact of fine-tuning on realistic human behavior simulation.

Evaluation Metrics We evaluate model performance across two key dimensions: **Next Action Generation** and **Session Outcome Classification**.

For **Next Action Generation**, we evaluate whether the model can accurately generate user actions during the process. We use an ***exact match accuracy***, where a generation is considered correct only if the action type, action target (e.g., search box or product link), and action attribute (such as search keyword) exactly match the ground truth. To avoid skewing the results toward longer sessions, we compute per-session accuracy first and then average across sessions, ensuring each session contributes equally to the final score.

To evaluate models' performance beyond human behavior simulation to shopping prediction, we introduce an additional task and metric focused on the **session final outcome**. The evaluation setting remains the same: the model is given the ground-truth session history up to the last step and tasked with generating the next (and final) action. Since the final action is always either a `click` action on the `buy now` button or a `terminate` action indicating closing the browser window, we evaluate the binary classification performance using the F1 score. This allows us to assess how well the model distinguishes between these two critical session outcomes (`buy` or `termination`).

4.3 Experimental Setup

We fine-tuned multiple language models to evaluate their performance on the action generation task. The models used in our experiments include:

- **Fine-Tuned Models:** Different versions of Llama 3.2, Qwen 2.5, and Mistral.
- **Baseline Models:** Different versions of Claude, Llama, Mistral, and DeepSeep-R1.

All fine-tuned models were trained using the same dataset and pipeline to ensure a fair comparison.

Model training was performed on a GPU cluster consisting of NVIDIA H200 GPUs, with each training job utilizing eight nodes, each equipped with eight GPUs, for a total of 64 GPUs, each with 140 GB of GPU memory. A typical job takes 3 hours on 64 GPUS, and in total, we used about 3700 H200 GPU hours for our experiments.

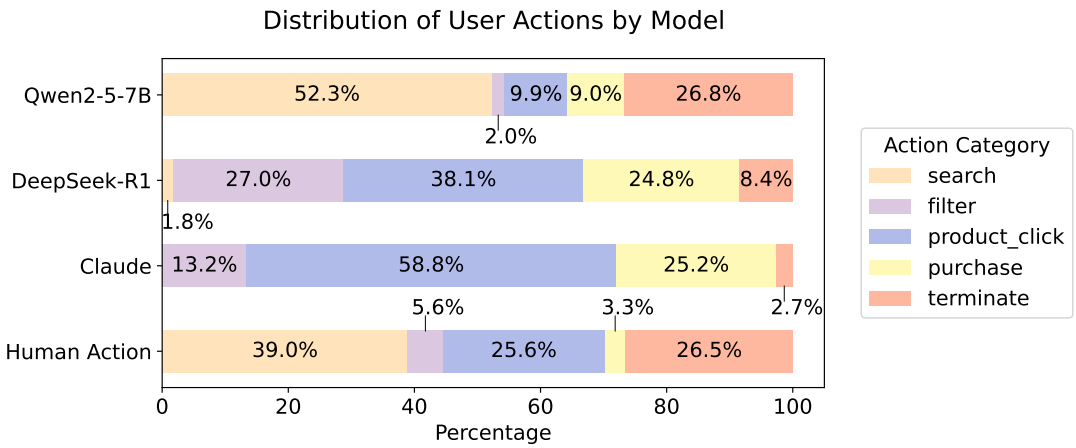

Figure 2: Action categories of human groundtruth, and generated by prompt-based Claude and by our fine-tuned models.

We employed Fully Sharded Data Parallel (Zhao et al., 2023) for efficient training. All sequences were padded or truncated to a context length of 40k tokens. We used a per-device batch size of 1, resulting in a global batch size of 64. The learning rate was set to 2e-5 with a cosine scheduler for adaptive learning rate adjustment. Models were trained for 1 epoch. For example, training Mistral-7B-v0.3 with a 40k token context window in our setup requires approximately 130GB of GPU memory per GPU, which is the largest model we can train.

4.4 Evaluation Results and Analysis

Following previous works (Lutz et al., 2024; Deng et al., 2023), we evaluate the performance of prompt-based LLMs in simulating human behavior on shopping scenarios, with results presented in Table 1. Our findings indicate that while state-of-the-art LLMs have demonstrated strong capabilities in various tasks and domains, *their ability to accurately simulate human behaviors remains limited*, with Claude 3.5 sonnet v2 achieving 11.69% accuracy in next action prediction. Compared to general instruction-tuned models, reasoning-focused DeepSeek-R1 achieved a higher accuracy of **11.86%** in the action generation task, suggesting that incorporating reasoning mechanisms offers some advantage. Similarly, for final outcome prediction, DeepSeek-R1 achieved an F1 score of **20.01%**, outperforming other baseline models. Additionally, larger model usually performs better than their smaller variants. These results indicate that reasoning ability positively impacts performance on human behavior prediction tasks, supporting the hypothesis that such tasks benefit from models trained with reasoning-oriented objectives.

We then fine-tuned various smaller open-sourced LLMs, including LLaMA 3.2, Qwen 2.5, and Mistral, using our training dataset. The results demonstrate that fine-tuning LLMs with action traces and synthesized reasoning traces significantly enhances performance. **Qwen 2.5-7B achieved 17.26% accuracy in action generation, significantly surpassing DeepSeek-R1 by 5.4% ($p < 10^{-10}$, McNemar's test).** Similarly, LLaMA 3.2-3B reached an F1 score of **33.99%** on the final outcome prediction task, further confirming the effectiveness of fine-tuning. Additionally, all fine-tuned models significantly outperformed their own ICL variants ($p < 10^{-5}$, McNemar's test). These findings underscore that incorporating domain-specific fine-tuning with synthesized reasoning traces leads to substantial improvements in the accuracy of human online shopping behavior simulation.

Figure 2 compares the distribution of actions generated by different models with real human behavior. Human users rarely apply filters and instead rely heavily on iterative search, averaging 2.82 searches per session, which is over seven times more frequent than filter actions. This reflects natural behaviors such as correcting typos and revising keywords (as shown in Sec. 4.6). In contrast, pre-trained LLMs such as Claude 3.5 Sonnet and DeepSeek-R1 tend to stick to the initial search keyword without revision, overuse filtering actions, and produce disproportionately high purchase rates that diverge from actual user behavior. We hypothesize that this bias arises because **existing LLM agent benchmarks like WebShop(Yao et al., 2022) and WebArena(Zhou et al., 2024) primarily evaluate task completion** (i.e., making a purchase), which incentivizes models to optimize for purchase-heavy trajectories rather than realistic browsing and search behavior.

The fine-tuned models exhibit a distribution that more closely aligns with human action patterns, capturing a more natural balance between search refinement, product clicks, and minimal reliance on filtering. Unlike pre-trained models, they do not prematurely converge on purchase-oriented behavior and instead reflect the exploratory and iterative nature of real user sessions.

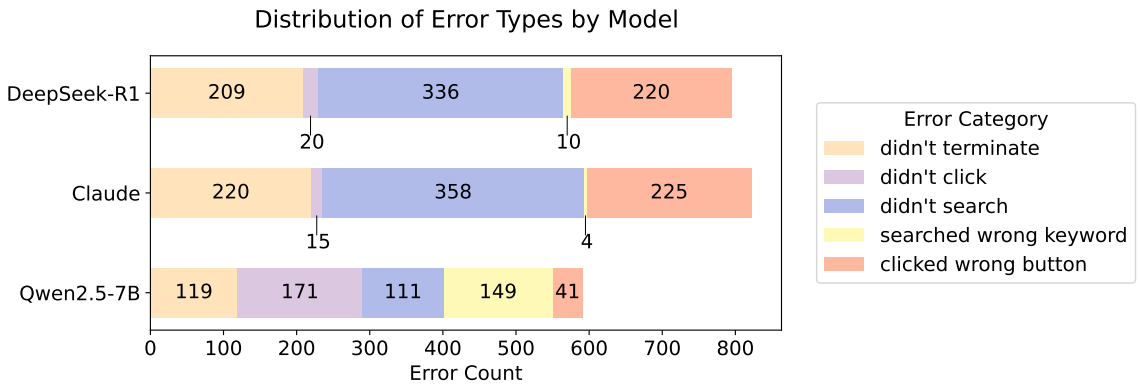

Figure 3: Error Type Analysis of different models.

	Example 1	**Example 2**
Previous Action	search for "disney gift"	search for "tee conector"
Human Next Action	search for "disney gift card"	search for "tee connector"
Qwen-2.5-7B	search for "disney gift card"	search for "tee connector"
Claude	click on `disney_gift_card_...`	click on `spalolen_30_pack_...`

Table 2: Model predictions vs. human next actions.

4.5 Ablation Study

To evaluate the impact of synthesized reasoning traces on the model's action generation capability, we conducted an ablation experiment. In the base setting, we removed reasoning traces from both the training and evaluation stages to isolate the contribution of reasoning-augmented learning. This setup allows us to directly assess whether exposure to synthesized reasoning traces improves the model's ability to generate user actions and predict outcomes.

From Table 1, most models exhibited substantial improvements in action generation accuracy when trained with reasoning traces, with relative gains ranging from 3.54% to 69.69%. Similar benefits were observed for final outcome prediction, where F1 scores increased by 25.78% to over 600% across models. For example, Qwen2.5-7B achieved a **33.86%** F1 score with reasoning traces but dropped to **26.92%** without them, underscoring the importance of explicit reasoning guidance. These results confirm that incorporating synthesized reasoning traces enhances both step-wise action generation and final outcome modeling, reinforcing their value in fine-tuning models for human behavior simulation.

4.6 Error Analysis

We analyze model behavior across different error types, comparing two ICL models, Claude, representing general-purpose chat models, and DeepSeek-R1, representing reasoning-augmented models, with a fine-tuned compact model, Qwen2.5–7B. We focus on five distinct error types. **Didn't terminate**, **didn't click**, and **didn't search** indicate cases where the model chose a different action instead of terminating, clicking, or searching, as the human user did. **Searched wrong keyword** refers to instances where the model performed a search like the user, but used a different (and incorrect) search query. **Clicked wrong button** captures cases where the model clicked, but on a different button than the one selected by the human user. Illegal actions generated by models are excluded from this analysis. Overall, Claude and DeepSeek-R1 exhibit very similar error profiles, suggesting that the reinforcement learning process used to incorporate reasoning introduces similar biases to those seen in standard chat-based ICL models. A major shared failure mode of ICL models is their tendency to continue sessions or make the purchase even when human users would have terminated them by closing the browser window. This aligns with earlier findings that ICL agents are more likely to complete purchases, likely because current LLMs are optimized to fulfill task goals rather than follow subtle social cues or termination heuristics. In contrast, fientuned Qwen2.5–7B demonstrates more accurate alignment with human termination behavior.

We also observe that Qwen-2.5–7B better captures iterative search behaviors commonly exhibited by real users, such as retrying a query with corrected keywords or fixing typos after an unsatisfactory result. In contrast, ICL-based models like Claude tend to persist with the original query result and proceed to a different type of action, rather than issuing a refined search. As shown in Table 2, Qwen-2.5–7B more closely matches human revisions in both examples, whereas Claude chooses to visit the product shown on the current search result page. These findings suggest that fine-tuning not only reduces the overall error rate but also enhances the model's ability to capture fine-grained user behavior patterns, such as correcting typos, retrying failed searches, and deciding when to terminate a session.

5 Discussion and Future Works

5.1 Action Misalign Between Human and LLM Agents

Existing research has shown that large language models (LLMs) can generate highly "believable" human behavior simulations, supporting various interactive and social simulation scenarios. However, our results indicate that general-purpose pretrained LLMs, despite their strong ability in a variety of tasks and applications, struggle to **accurately** generate user actions. This is reflected in their performance—DeepSeek-R1 and Claude 3.5 Sonnet v2, for example, achieved only **11.86%** and **11.69%** accuracy, respectively, on the next action generation task. These findings suggest that while LLMs may appear human-like and produce plausible and believable behaviors, accurate step-wise next action prediction requires additional fine-tuning and explicit alignment with real human actions.

Building on this observation, we find that current LLM agents are often misaligned with real human users, particularly in the online shopping domain. For example, as shown in Figure 2 and Figure 3, out-of-the-box LLMs tend to make more purchases than human users. We hypothesize that this stems from the training objectives of many LLMs, which are typically optimized for the final task completion (e.g., sending emails or making purchases) and evaluated based on task completion and task efficiency (i.e., completing the task with a minimal number of steps). To enable accurate human behavior simulation, it is essential to close this gap and ensure models capture what a real user will do under certain circumstances.

We also advocate for the development of evaluation metrics that reflect the irrational nature of human behavior. On one hand, many user actions are rational and goal-directed, where a single correct next step exists and can be evaluated against a gold standard. On the other hand, users also exhibit irrational or weakly deterministic behaviors, such as semi-random clicks, and typographical errors that even the same person may not replicate consistently. A robust metric should not penalize plausible variants that remain consistent with human trajectories and should explicitly assess an agent's ability to reproduce human-like error patterns. For example, the capacity to generate *human-like typos* is also essential for applications like automatic UI/UX evaluation (Lu et al., 2025b), where assessing typo correction features requires agents to reproduce the kinds of errors real users make. A human-aligned evaluation should therefore distinguish true errors from acceptable variability and report performance across both rational and non-rational action regimes.

5.2 Reasoning in Next Action Generation

Our experiments further highlight the critical role of synthesized reasoning traces in improving action generation. Removing reasoning traces from the training data led to a notable performance drop across all models. Additionally, most models showed measurable improvements when trained with reasoning traces, reinforcing the importance of providing explicit reasoning signals. These results suggest that reasoning traces not only enhance model interpretability but also serve as a guiding mechanism, enabling the model to make more contextually appropriate and human-aligned decisions.

Overall, our findings emphasize the necessity of domain-specific fine-tuning using enriched, context-aware human behavioral data. By incorporating synthesized reasoning traces, we bridge the gap between the believable yet imprecise predictions of general LLMs and the accurate behavioral simulations required for real-world applications.

5.3 Future Works

Future research should move beyond treating reasoning traces as static supervision and instead focus on actively enhancing the model's reasoning abilities. One direction is to use reinforcement learning to improve the reasoning for action prediction accuracy, where better reasoning is defined by its ability to produce more accurate action prediction. Another direction is to train models to generate reasoning that more closely reflects human cognitive processes, so the reasoning itself can serve as a source of qualitative insight when analyzing or interacting with LLM agents. In addition to advancing reasoning, future work could scale up the dataset to improve robustness, incorporate user personas for more personalized simulations, and incorporate user intents with intent detection methods. Finally, incorporating vision-language models (VLMs) may enhance the model's understanding of graphical interfaces in complex web environments.

6 Conclusion

In this work, we present the first quantitative, process-centric evaluation of state-of-the-art large language models (LLMs) for simulating human behavior in an online shopping task. Our study demonstrates that LLMs fine-tuned with real-world human behavioral data and synthesized reasoning traces showed a significant enhancement in their ability to generate user actions. Compared to in-context learning and pretrained baselines, our method achieved substantial improvements, including a **13.98%** absolute gain in F1 score for final outcome prediction and a **5.4%** increase in action generation accuracy.

These results underscore the critical role of explicit reasoning traces in aligning model predictions with human behavior. By enriching behavioral datasets with structured reasoning, we move closer to accurate and interpretable simulations of human behavior. Our findings highlight a promising direction for the development of LLM agent systems capable of producing realistic and explainable human-like behaviors in online shopping and other interactive domains.

Limitations

Our study has several limitations that should be considered when interpreting the results. First, following recent works (DeepSeek-AI et al., 2025), we only evaluated the extent to which synthetic reasoning traces improve model performance. Conducting human evaluations on the interpretability and usefulness of the generated reasoning traces could better assess how well these traces support human comprehension and trust in the model's predictions. Second, our experiments were conducted exclusively on a single behavioral task (online shopping) and the generalizability of our findings to other tasks and contexts remains uncertain. Third, we have not yet evaluated the model on real human-annotated datasets containing authentic reasoning traces, making it unclear how well the synthesized reasoning trace aligns with human reasoning. Additionally, the process of generating synthesized reasoning traces may introduce unintended biases, potentially impacting prediction accuracy and interoperability. Finally, to simplify the experimental setup, we limited the action space to basic browser operations such as `type` and `click`. Incorporating more complex interactions, such as scrolling, waiting, or hover actions, would allow for a more realistic simulation of human behavior in web environments and offer deeper insights into how LLMs handle nuanced browser-based simulation.

References

Chaoran Chen, Bingsheng Yao, Ruishi Zou, Wenyue Hua, Weimin Lyu, 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. *Preprint*, arXiv:2502.13012.

Jiaju Chen, Yuxuan Lu, Xiaojie Wang, Huimin Zeng, Jing Huang, Jiri Gesi, Ying Xu, Bingsheng Yao, and Dakuo Wang. 2025b. Multi-Agent-as-Judge: Aligning LLM-Agent-Based Automated Evaluation with Multi-Dimensional Human Evaluation. *Preprint*, arXiv:2507.21028.

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, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2Web: Towards a Generalist Agent for the Web. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Vardhan Dongre, Xiaocheng Yang, Emre Can Acikgoz, Suvodip Dey, Gokhan Tur, and Dilek Hakkani-Tür. 2025. ReSpAct: Harmonizing Reasoning, Speaking, and Acting Towards Building Large Language Model-Based Conversational AI Agents. *Preprint*, arXiv:2411.00927.

David W. Eccles and Güler Arsal. 2017. The think aloud method: What is it and how do I use it? *Qualitative Research in Sport, Exercise and Health*, 9(4):514–531.

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*.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel

Fried. 2024. VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 881–905, Bangkok, Thailand. Association for Computational Linguistics.

Yuxuan Lu, Bingsheng Yao, Hansu Gu, Jing Huang, Jessie Wang, Laurence Li, Jiri Gesi, Qi He, Toby Jia-Jun Li, and Dakuo Wang. 2025a. UXAgent: An LLM Agent-Based Usability Testing Framework for Web Design. *Preprint*, arXiv:2502.12561.

Yuxuan Lu, Bingsheng Yao, Hansu Gu, Jing Huang, Jessie Wang, Yang Li, Jiri Gesi, Qi He, Toby Jia-Jun Li, and Dakuo Wang. 2025b. UXAgent: A System for Simulating Usability Testing of Web Design with LLM Agents. *Preprint*, arXiv:2504.09407.

Qinyu Luo, Yining Ye, Shihao Liang, Zhong Zhang, Yujia Qin, Yaxi Lu, Yesai Wu, Xin Cong, Yankai Lin, Yingli Zhang, Xiaoyin Che, Zhiyuan Liu, and Maosong Sun. 2024. RepoAgent: An LLM-Powered Open-Source Framework for Repository-level Code Documentation Generation. *Preprint*, arXiv:2402.16667.

Michael Lutz, Arth Bohra, Manvel Saroyan, Artem Harutyunyan, and Giovanni Campagna. 2024. WILBUR: Adaptive In-Context Learning for Robust and Accurate Web Agents. *Preprint*, arXiv:2404.05902.

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*, UIST '23, pages 1–22, New York, NY, USA. Association for Computing Machinery.

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.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative Agents for Software Development. *Preprint*, arXiv:2307.07924.

Lu Sun, Shihan Fu, Bingsheng Yao, Yuxuan Lu, Wenbo Li, Hansu Gu, Jiri Gesi, Jing Huang, Chen Luo, and Dakuo Wang. 2025. LLM Agent Meets Agentic AI: Can LLM Agents Simulate Customers to Evaluate Agentic-AI-based Shopping Assistants? *Preprint*, arXiv:2509.21501.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Preprint*, arXiv:2201.11903.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can Large Language Model Agents Simulate Human Trust Behavior? *Preprint*, arXiv:2402.04559.

Shunyu Yao, Howard Chen, John Yang, and Karthik R. Narasimhan. 2022. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. In *Advances in Neural Information Processing Systems*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. *Preprint*, arXiv:2210.03629.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel. *Preprint*, arXiv:2304.11277.

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. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. *Preprint*, arXiv:2307.13854.

A Prompts

Reasoning Synthesize Prompt:

```
You will be given a customer's shopping
↪ journey on one of the largest e-commerce
↪ platforms globally. you will be given
↪ the context (what the user is looking
↪ at), the action (what the user did), and
↪ your job is to predict the user's
↪ rationale for the action. The rationale
↪ should follow
Here is an example:
{example}
For each action in the input, output a
↪ rationale.
If the action is "terminate", it means that
↪ you didn't find any desired product and
↪ you decided to leave the website by
↪ closing the browser window.
```

Baseline model evaluation prompt:

```
<IMPORTANT>
Your task is to predict the next action and
↪ provide rationale for the action based
↪ on the previous actions and context.
You need to pretend that you are a user,
↪ browsing one of the largest e-commerce
↪ platforms globally and searching for a
↪ product to purchase.
The history action (with details described
↪ below) and context will be provided to
↪ you.
You need to predict the next action and
↪ provide rationale for the action.
</IMPORTANT>


# Action Space

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

#### 1. `type_and_submit`:
Type text into an input field and
↪ immediately submit the form. Equivalent
↪ to typing text into an input and
↪ pressing enter key.

{
    "type": "type_and_submit",
    "name": "input_name",
    "text": "search_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"
}

# Context
Your context will be an **simplified
↪ version** of the raw HTML of the one of
↪ the largest e-commerce platforms
↪ globally 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 type_and_submit).

# Rationale

The rationale is a first-person sentence of
↪ what you are thinking when you make the
↪ action. It should be a short sentence
↪ that explains why you are making the
↪ action.

# Output Format

You need to predict the next action and
↪ provide rationale for the action. Your
↪ output should follow a strict JSON form:

{
    "action": {
        // action goes here
        "type": "<type>",
        ...
    },
    "rationale": "<rationale>" // rationale
    ↪ goes here, a string
}

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

Model	Generated Action			Session Outcome		
	Accuracy	%Δ vs Base	v.s. DS-R1	F1 Score	%Δ vs Base	v.s. DS-R1
Open-Source Models						
DeepSeek-R1	11.86%	-	-	20.01%	-	-
Llama 3.1 8B	5.05%	-	-6.81%	10.87%	-	-9.14%
Llama 3.1 70B	8.19%	-	-3.67%	12.69%	-	-7.32%
Mixtral 8x7B	5.41%	-	-6.45%	13.16%	-	-6.85%
Qwen2.5-7B	4.25%	-	-7.61%	11.94%	-	-8.07%
Qwen2.5-3B	3.91%	-	-7.95%	10.87%	-	-9.14%
Qwen2.5-1.5B	3.27%	-	-8.59%	7.94%	-	-12.07%
Mistral-7B-v0.3	4.25%	-	-7.61%	11.27%	-	-8.74%
Llama 3.2 3B	2.93%	-	-8.93%	8.60%	-	-11.41%
Llama 3.2 1B	3.71%	-	-8.15%	3.09%	-	-16.92%
Proprietary Models						
Claude 3.5 Haiku	9.18%	-	-2.68%	14.77%	-	-5.24%
Claude 3 Opus	6.78%	-	-5.08%	15.08%	-	-4.93%
Claude 3 Sonnet	8.42%	-	-3.44%	17.40%	-	-2.61%
Claude 3.5 Sonnet	9.72%	-	-2.14%	15.91%	-	-4.10%
Claude 3.5 Sonnet v2	11.69%	-	-0.17%	18.54%	-	-1.47%
Claude 3.7 Sonnet	9.34%	-	-2.52%	12.81%	-	-7.20%
Fine-tuned Models						
Qwen2.5-7B	16.67%	-	4.81%	26.92%	-	6.91%
+ reasoning	**17.26%**	3.54%	5.40%	33.86%	25.78%	13.85%
Qwen2.5-3B	14.53%	-	2.67%	22.88%	-	2.87%
+ reasoning	11.88%	-18.24%	0.02%	28.52%	24.65%	8.51%
Qwen2.5-1.5B	5.03%	-	-6.83%	5.67%	-	-14.34%
+ reasoning	16.06%	219.28%	4.20%	27.69%	388.36%	7.68%
Mistral-7B-v0.3	14.17%	-	2.31%	17.99%	-	-2.02%
+ reasoning	15.84%	11.79%	3.98%	30.12%	67.43%	10.11%
Llama-3.2-3B	9.31%	-	-2.55%	4.73%	-	-15.28%
+ reasoning	15.77%	69.39%	3.91%	**33.99%**	618.60%	13.98%
Llama-3.2-1B	11.13%	-	-0.73%	10.44%	-	-9.57%
+ reasoning	7.53%	-32.35%	-4.33%	15.08%	44.44%	-4.93%

Table 3: Performance comparison of models in different settings.

B Example Context

```
<html>

<head>
    <title>Amazon.com : pressure
    ↪  washer</title>
</head>

<body>
    <form role="search"> <input
    ↪  name="search_input" value="pressure
    ↪  washer" type="text"
    ↪  aria-label="Search Amazon" />
        <input name="search_button"
        ↪  type="submit" value="Go" />
    </form>
    <div name="refinements">
       <div> <span class="refinement-title">
        ↪  Amazon Prime </span>
            <li name="refinements.amazon_⌋
            ↪  prime.prime_eligible"
            ↪  role="checkbox"> Prime
            ↪  Eligible <input
            ↪  type="checkbox">
            </li>
        </div>
       <div> <span class="refinement-title">
        ↪  Prime Delivery </span>
            <li name="refinements.prime_⌋
            ↪  delivery.tomorrow_by_8am"
            ↪  role="checkbox"> Tomorrow
            ↪  by 8AM <input
                    type="checkbox"> </li>
        </div>
       <div> <span class="refinement-title">
        ↪  Delivery Day </span>
            <li name="refinements.delivery_⌋
            ↪  day.get_it_today"
            ↪  role="checkbox"> Get It
            ↪  Today <input
            ↪  type="checkbox"> </li>
            <li name="refinements.delivery_⌋
            ↪  day.get_it_by_tomorrow"
            ↪  role="checkbox"> Get It by
            ↪  Tomorrow <input
                    type="checkbox"> </li>
        </div>
       <div> <span class="refinement-title">
        ↪  Customer Reviews </span>
            <li name="refinements.customer_⌋
            ↪  reviews.4_stars__up"
            ↪  role="checkbox"> 4 Stars
            ↪  &amp; Up <input
                    type="checkbox"> </li>
        </div>
       <div> <span class="refinement-title">
        ↪  All Top Brands </span>
            <li name="refinements.all_top_⌋
            ↪  brands.top_brands"
            ↪  role="checkbox"> Top Brands
            ↪  <input type="checkbox">
            ↪  </li>
    </div>
   <div> <span class="refinement-title">
    ↪  Deals &amp; Discounts </span>
            <li name="refinements.deals__⌋
            ↪  discounts.all_discounts"
            ↪  role="checkbox"> All
            ↪  Discounts <input
            ↪  type="checkbox">
            </li>
            <li name="refinements.deals__⌋
            ↪  discounts.todays_deals"
            ↪  role="checkbox">
            ↪  Today&#39;s Deals <input
                    type="checkbox"> </li>
    </div>
   <div> <span class="refinement-title">
    ↪  Pressure Washer Pressure
    ↪  </span>
            <li name="refinements.pressure_⌋
            ↪  washer_pressure.under_⌋
            ↪  1700_psi" role="checkbox">
            ↪  Under 1700 PSI <input
                    type="checkbox">  </li>
            <li name="refinements.pressure_⌋
            ↪  washer_pressure.1700_to_⌋
            ↪  1999_psi" role="checkbox">
            ↪  1700 to 1999 PSI <input
                    type="checkbox">  </li>
            <li name="refinements.pressure_⌋
            ↪  washer_pressure.2000_to_⌋
            ↪  2599_psi" role="checkbox">
            ↪  2000 to 2599 PSI <input
                    type="checkbox">  </li>
            <li name="refinements.pressure_⌋
            ↪  washer_pressure.2600_to_⌋
            ↪  2799_psi" role="checkbox">
            ↪  2600 to 2799 PSI <input
                    type="checkbox">  </li>
            <li name="refinements.pressure_⌋
            ↪  washer_pressure.2800_to_⌋
            ↪  3099_psi" role="checkbox">
            ↪  2800 to 3099 PSI <input
                    type="checkbox">  </li>
            <li name="refinements.pressure_⌋
            ↪  washer_pressure.3100_to_⌋
            ↪  3999_psi" role="checkbox">
            ↪  3100 to 3999 PSI <input
                    type="checkbox">  </li>
            <li name="refinements.pressure_⌋
            ↪  washer_pressure.4000_psi__⌋
            ↪  above" role="checkbox">
            ↪  4000 PSI &amp; Above <input
                    type="checkbox">  </li>
    </div>
```

```
<div> <span class="refinement-title">
 ↪  Condition </span>
     <li
      ↪  name="refinements.condition.renewed"
      ↪  role="checkbox"> Renewed
      ↪  <input type="checkbox">
      ↪  </li>
 <li name="refinements.condition.new"
      ↪  role="checkbox"> New <input
      ↪  type="checkbox"> </li>
     <li
      ↪  name="refinements.condition.used"
      ↪  role="checkbox"> Used <input
      ↪  type="checkbox"> </li>
 </div>
 ...
<div> <span class="refinement-title">
 ↪  Color </span>
   <li name="refinements.color.black"
      ↪  role="checkbox"> Black
      ↪  <input type="checkbox">
      ↪  </li>
   <li name="refinements.color.multi"
      ↪  role="checkbox"> Multi
      ↪  <input type="checkbox">
      ↪  </li>
    <li name="refinements.color.blue"
      ↪  role="checkbox"> Blue <input
      ↪  type="checkbox"> </li>
    <li name="refinements.color.grey"
      ↪  role="checkbox"> Grey <input
      ↪  type="checkbox"> </li>
   <li name="refinements.color.white"
      ↪  role="checkbox"> White
      ↪  <input type="checkbox">
      ↪  </li>
   <li name="refinements.color.brown"
      ↪  role="checkbox"> Brown
      ↪  <input type="checkbox">
      ↪  </li>
   <li name="refinements.color.beige"
      ↪  role="checkbox"> Beige
      ↪  <input type="checkbox">
      ↪  </li>
     <li name="refinements.color.red"
      ↪  role="checkbox"> Red <input
      ↪  type="checkbox"> </li>
    <li name="refinements.color.pink"
      ↪  role="checkbox"> Pink <input
      ↪  type="checkbox"> </li>
  <li name="refinements.color.orange"
      ↪  role="checkbox"> Orange
      ↪  <input type="checkbox">
      ↪  </li>
  <li name="refinements.color.yellow"
      ↪  role="checkbox"> Yellow
      ↪  <input type="checkbox">
      ↪  </li>
   <li name="refinements.color.ivory"
      ↪  role="checkbox"> Ivory
      ↪  <input type="checkbox">
      ↪  </li>
   <li name="refinements.color.green"
      ↪  role="checkbox"> Green
      ↪  <input type="checkbox">
      ↪  </li>
  <li name="refinements.color.purple"
      ↪  role="checkbox"> Purple
      ↪  <input type="checkbox">
      ↪  </li>
   <li name="refinements.color.gold"
      ↪  role="checkbox"> Gold <input
      ↪  type="checkbox"> </li>
  <li name="refinements.color.silver"
      ↪  role="checkbox"> Silver
      ↪  <input type="checkbox">
      ↪  </li>
   <li name="refinements.color.clear"
      ↪  role="checkbox"> Clear
      ↪  <input type="checkbox">
      ↪  </li>
 </div>
<div> <span class="refinement-title">
 ↪  More-sustainable Products
 ↪  </span>
     <li
      ↪  name="refinements.moresustainable_⌋
      ↪  products.climate_pledge_⌋
      ↪  friendly" role="checkbox">
      ↪  Climate Pledge
         Friendly <input
          ↪  type="checkbox"> </li>
 </div>
</div>
<div>
 <div class="search-result"> <a
         name="search_results.2025_⌋
         ↪  upgraded_electric_⌋
         ↪  pressure_washer_with_⌋
         ↪  adjustable_touch_⌋
         ↪  screen_5000_psi_33_⌋
         ↪  gpm_8_power_settings_⌋
         ↪  for_car_patio__floor_⌋
         ↪  cleaning_detergent_⌋
         ↪  tank_with_4_nozzles_⌋
         ↪  23ft_hose_35ft_⌋
         ↪  cord.view_product"
         class="product-name"> 2025
         ↪  Upgraded Electric
         ↪  Pressure Washer with
         ↪  Adjustable Touch Screen,
         ↪  5000 PSI
         3.3
```

```
        GPM, 8 Power Settings, for
        ↪ Car, Patio &amp; Floor
        ↪ Cleaning, Detergent
        ↪ Tank, with 4 Nozzles,
        ↪ 23ft
        Hose,
        35ft Cord </a>
    <div class="product-review">
    ↪ <span
    ↪ class="product-rating">5.0
    ↪ out of 5 stars</span><span
      class="product-rating-count">
          ↪ 20.0 reviews
          ↪ </span> </div>
    <div class="product-price"><┘
    ↪ span>$199.99</span></div>
  <div class="product-delivery">FREE
    ↪ delivery Mon Mar 17</div>
</div>
<div class="search-result"> <a
        name="search_┘
        ↪ results.2025upgraded_┘
        ↪ electric_pressure_┘
        ↪ washer_4500_psi_32_┘
        ↪ gpm_power_washer_with_┘
        ↪ 4_quick_connect_┘
        ↪ nozzles_inlet_hose__┘
        ↪ filter_foam_cannon_┘
        ↪ for_┘
        ↪ carsfencesdrivewayshome_┘
        ↪ cleaning.view_product"
        class="product-name">
        ↪ 2025Upgraded Electric
        ↪ Pressure Washer, 4500
        ↪ PSI 3.2 GPM Power Washer
        ↪ with 4
        Quick
        Connect Nozzles, Inlet Hose
        ↪ &amp; Filter&amp; Foam
        ↪ Cannon for
        ↪ Cars/Fences/Driveways/Home
        ↪ Cleaning
    </a>
    <div class="product-review">
    ↪ <span
    ↪ class="product-rating">5.0
    ↪ out of 5 stars</span><span
      class="product-rating-count">
          ↪ 19.0 reviews
          ↪ </span> </div>
    <div class="product-price">
    ↪ span>$159.99</span></div>
  <div class="product-delivery">FREE
    ↪ delivery Mon Mar 17</div>
</div>
<div class="search-result"> <a
      name="search_results.kärcher_┘
        ↪ pressure_washer_┘
        ↪ k1800ps_max_2250_psi_┘
        ↪ 3_spray_nozzles_┘
        ↪ detergent_tank_for_┘
        ↪ cars_driveways_siding_┘
        ↪ patios_146_max_┘
        ↪ gpm.view_product"
      class="product-name"> Kärcher
        ↪ Pressure Washer K1800PS,
        ↪ Max 2250 PSI, 3 Spray
        ↪ Nozzles, Detergent
      Tank,
      For Cars, Driveways, Siding,
        ↪ Patios, 1.46 max. GPM
        ↪ </a>
    <div class="product-review">
    ↪ <span
    ↪ class="product-rating">4.3
    ↪ out of 5 stars</span><span
      class="product-rating-count">
          ↪ 8520.0 reviews
          ↪ </span> </div>
    <div class="product-price">
    ↪ span>$167.99</span></div>
</div>
<div class="search-result"> <a
      name="search_results.kärcher_┘
        ↪ pressure_washer_k1700_┘
        ↪ max_2125_psi_3_spray_┘
        ↪ nozzles_detergent_┘
        ↪ tank_for_cars_┘
        ↪ driveways_siding_┘
        ↪ patios_146_max_┘
        ↪ gpm.view_product"
      class="product-name"> Kärcher
        ↪ Pressure Washer K1700,
        ↪ Max 2125 PSI, 3 Spray
        ↪ Nozzles, Detergent Tank,
      For Cars, Driveways, Siding,
        ↪ Patios, 1.46 max. GPM
        ↪ </a>
    <div class="product-review">
    ↪ <span
    ↪ class="product-rating">4.3
    ↪ out of 5 stars</span><span
      class="product-rating-count">
          ↪ 8520.0 reviews
          ↪ </span> </div>
    <div class="product-price">
    ↪ span>$145.99</span></div>
</div>
<div class="search-result"> <a
```

```
            name="search_results.kärcher_⌋
              ↪  pressure_washer_⌋
              ↪  k2300ps_max_2875_psi_⌋
              ↪  4_spray_nozzles_⌋
              ↪  detergent_tank_hose_⌋
              ↪  reel_for_cars_⌋
              ↪  driveways_siding_⌋
              ↪  patios_207_max_⌋
              ↪  gpm.view_product"
            class="product-name"> Kärcher
              ↪  Pressure Washer K2300PS,
              ↪  Max 2875 PSI, 4 Spray
              ↪  Nozzles, Detergent
             Tank, Hose Reel, For Cars,
              ↪  Driveways, Siding,
              ↪  Patios, 2.07 max. GPM
              ↪  </a>
          <div class="product-review">
          ↪  <span
          ↪  class="product-rating">4.3
          ↪  out of 5 stars</span><span
            class="product-rating-count">
                ↪  8520.0 reviews
                ↪  </span> </div>
          <div class="product-price"><⌋
          ↪  span>$223.98</span></div>
      </div>
      <span name="pagination">
          <span name="pagination.1"
          ↪  aria-label="Current page,
          ↪  page 1">1</span>
          <a name="pagination.2"
          ↪  aria-label="Go to page
          ↪  2">2</a>
          <a name="pagination.3"
          ↪  aria-label="Go to page
          ↪  3">3</a>
          <a aria-label="...">...</a>
          <a name="pagination.7"
          ↪  aria-label="Go to page
          ↪  7">7</a>
      </span>
  </div>
</body>

</html>
```