Orion: Steering Personalized Web Agents via Global-Micro Profiling and Adaptive Intent Tracking

Die Hu[1,2], Jingguo Ge[1,2*], Weitao Tang[2], He Kong[1,2], Liangxiong Li[1], Bingzhen Wu[1]

[1]State Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{hudie, gejingguo, tangweitao, konghe, liliangxiong, wubingzhen}@iie.ac.cn

Abstract

Recently, Large Language Models (LLMs) based web agents have shown significant potential in web understanding and interaction tasks. However, their personalization ability and user experience remain limited by the ambiguity and dynamic nature of user intent, struggling to model diverse user interests and track intent changes over time. To address these challenges, this paper proposes Orion, a novel personalized web agent. Orion adopts a global-micro profiling mechanism to balance users' long-term stable preferences and scenario-based needs, and introduces context-aware interest retrieval to enhance personalization. Additionally, we design adaptive profile tracking and proactive disambiguation mechanisms to effectively address the continuous evolution of user intent in multi-turn interactions. Orion is optimized through end-to-end online reinforcement learning, improving personalized reasoning and decision-making ability in real interactive scenarios. Experiments demonstrate that Orion significantly outperforms state-of-the-art baselines in personalized understanding and task efficiency.

1 Introduction

LLM-driven web agents show great promise in promoting the intelligent process of human-computer interaction (Wang et al. 2024; Xi et al. 2025). Despite this promise, existing web agents (Deng et al. 2023; Zhou et al. 2023) generally adopt a single-turn "instruction–response" paradigm (see Figure 1(a)), relying solely on users' initial instructions to drive task execution. However, in real-world applications, user instructions are often hard to directly and fully express personalized preferences, especially in highly personalized scenarios such as e-commerce, where user satisfaction is affected by multiple factors such as price and brand. At the same time, deep interests and real intentions are difficult to explicitly reflect through instructions (Stengel-Eskin et al. 2023; Liu et al. 2023). This limitation seriously restricts the personalization ability and user experience of existing web agents. Therefore, it is urgent to build personalized web agents that can truly understand users.

However, there are still several main challenges in personalized web agents:

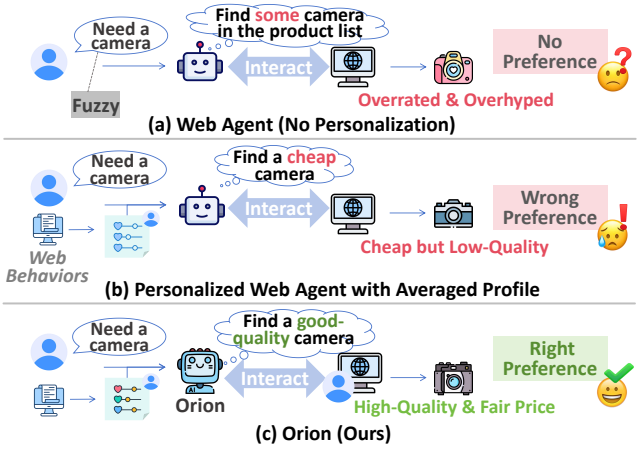

Figure 1: The motivation of our Orion.

First, there are **inherent limitations in user intent expression**. In actual web interactions, it is difficult for users to fully express their needs through a single initial instruction, resulting in incomplete and ambiguous semantics. Although existing methods (Cai et al. 2024; Shi et al. 2025) attempt to aggregate long-term historical interactions to assist understanding, they typically average user behaviors into a static profile, significantly compressing interest diversity and ignoring context specificity. For example, as shown in Figure 1(b), a user might usually prefer cost-effective daily necessities, but also choose high-end digital products in specific situations. The averaged profile can only capture the general preference for cost-effectiveness, but fails to reflect fine-grained and context-related real intentions. Moreover, excessive reliance on full historical context increases computational overhead and can dilute attention, thus reducing reasoning efficiency and accuracy.

Second, the **dynamic evolution of user intent and preferences** during multi-turn web interactions presents another key challenge (Liu et al. 2023). User intent is inherently dynamic, continuously adjusting in response to agent feedback, page content, and the user's own changing considerations. The limited and ambiguous nature of initial instructions often leads the agent to early misinterpretations, which accu-

*Corresponding author

mulate and amplify over subsequent interactions, ultimately causing the agent to deviate further from the user's actual needs, as illustrated in Figure 1(a), 1(b)). This compound effect of dynamic intent evolution and error propagation limits the agent's ability to accurately track the latest user intent, thereby hindering further advancement in personalized experience.

To address these challenges, we propose **Orion**, a personalized web agent framework, as illustrated in the simplified diagram in Figure 1(c). This framework jointly models users' long-term stable preferences and scenario-specific fine-grained interests. Through a context-aware profile retrieval mechanism, it effectively avoids interest expression dilution and retrieval efficiency bottlenecks. At the same time, the system introduces adaptive profile tracking and proactive disambiguation methods to achieve dynamic updating and real-time adaptation of user intent during multi-turn interactions. Finally, through end-to-end online reinforcement learning (RL), leveraging Supervised Fine-Tuning (SFT) (Foster, Block, and Misra 2024) and the Group Relative Policy Optimization (GRPO) algorithm (Guo et al. 2025), Orion significantly improves personalized reasoning and decision-making ability. The main contributions of this paper are as follows:

- We propose a novel global-micro profile method that efficiently models users' long-term preferences and fine-grained contextual interests, enhancing web agents' for personalized understanding capacity.
- We design adaptive profile tracking and proactive disambiguation mechanisms, significantly enhancing the agent's intent tracking and adaptability throughout multi-turn interactions;
- We compare Orion with state-of-the-art methods across 3 personalized web tasks. Orion achieves an accuracy improvement of 17.2% to 20.0% over the strongest baselines and demonstrates superior capability in dynamic, multi-turn interactions.

2 Related Work

Web Agent LLM-based web agents are designed to autonomously interact with web environments and accomplish tasks following user instructions(Nguyen et al. 2025; Ning et al. 2025). Early studies mostly focused on single-turn interaction, typically conducted in offline or simplified simulation environments. MiniWoB++ (Humphreys et al. 2022) defined the basic interaction paradigm based on web components. WebShop (Yao et al. 2022) and Mind2Web (Deng et al. 2023) further enhanced task realism and complexity by simulating shopping websites and incorporating large-scale real web page screenshots, respectively. More recent work has moved toward complex, multi-turn interactions in online environments. WebArena (Zhou et al. 2023) constructed an independent environment with multiple realistic websites, enabling agents to perform complex cross-application workflows. ChatShop (Chen, Wiseman, and Dhingra 2024) focuses on conversational interaction, enabling agents to negotiate with users through multi-turn dialogs to complete shopping tasks.

Despite advances in task execution, these web agents still struggle with understanding and adapting to user preferences. They are unable to provide customized services based on individual histories, preferences, or habits. This one-size-fits-all paradigm limits both service quality and user satisfaction.

Personalized Web Agent To address personalization challenges, Cai et al. (Cai et al. 2024) proposed the PUMA framework, which constructs a memory bank for each user based on their entire interaction history and supports personalized function calls via task-specific retrieval strategies. However, relying on static, holistic user profiles makes it hard to capture the evolution of user preferences and respond to immediate, contextual intentions. Furthermore, its memory retrieval mechanism may encounter efficiency bottlenecks when scaling to extensive user histories.

3 Problem Formulation

Current web agents mainly adopt two types of interaction paradigms: GUI-based and API-based. GUI-based methods are centered on human visual perception, mapping high-level semantic tasks into low-level visual action sequences (Wang et al. 2024). Such human-oriented design is not optimal for computer-based interaction. In contrast, API-based methods leverage structured state feedback and high-level action abstraction, enabling agents to focus more effectively on task decisions and personalized reasoning. In this work, we focus on API-based web agents and formally model the multi-turn, dynamic, and partially observable human-computer interaction process as a Partially Observable Markov Decision Process (POMDP) (Kurniawati 2021).

In each interaction episode, the user provides a global goal instruction I at the initial time ($t = 0$). The agent can access the user's historical behavior archive $H = \{h_1, h_2, \ldots, h_{|H|}\}$, where each h_i contains multimodal information such as product title, category, brand, review text, and timestamp. At each step t, the agent outputs an action a_t from the action space $\mathcal{A}$. The environment returns an observation o_t. The entire interaction process forms a trajectory $\tau = \{(a_1, o_1), \ldots, (a_T, o_T)\}$. The agent policy π aims to generate an action sequence to maximize the cumulative reward $\sum_{t=1}^{T} R(a_t, o_t)$, where $R(a_t, o_t)$ measures the overall contribution of the current action to task completion and personalized requirement satisfaction.

4 Methodology

The architecture of Orion is illustrated in Figure 2. Our Orion begins with **1) Global-Micro User Profile Construction**, where LLMs are employed to construct a fine-grained understanding of user preferences, encompassing both long-term stable attributes and context-specific needs. Next, we utilize a context-aware profile retrieval module to extract the most relevant preferences based on the user's instruction. This retrieved context then informs the **2) Adaptive Agent Decision-Making** process, which is designed to track evolving user intents through a dynamically updated contextual

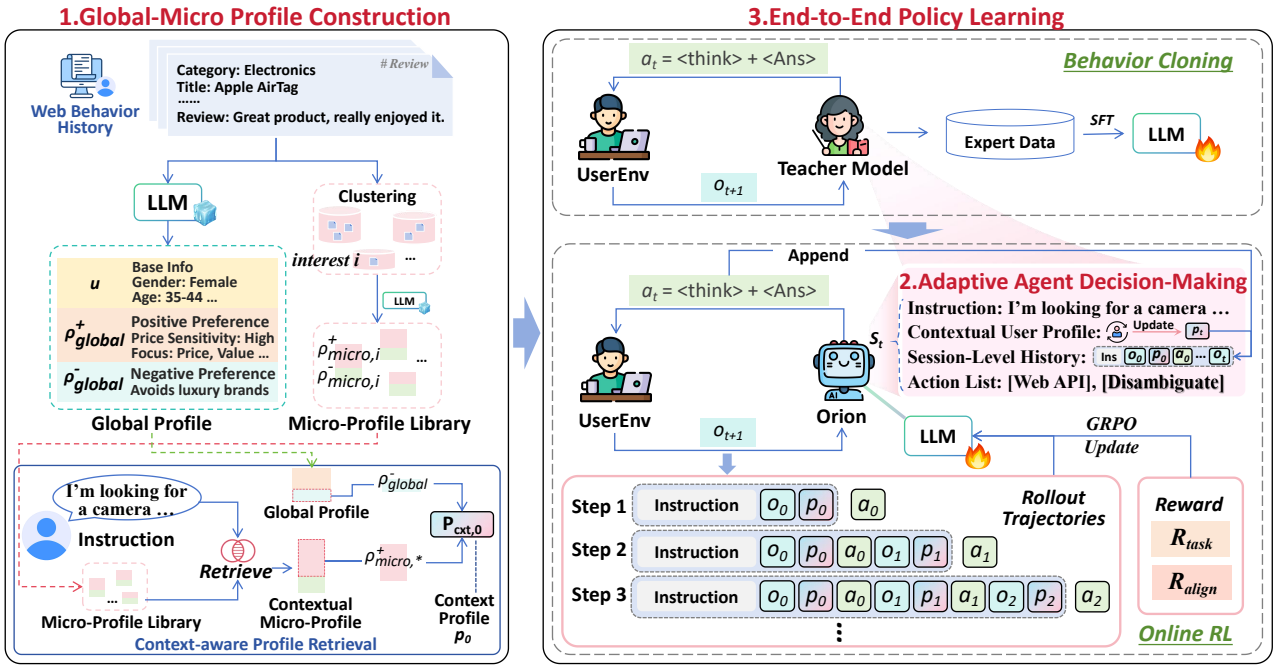

Figure 2: The overall architecture of our Orion.

profile and an expanded action space. Finally, Orion's interactive policy is optimized through **3) End-to-End Policy Learning**, employing behavior cloning for foundational reasoning, followed by online RL for continuous optimization. This design empowers Orion to efficiently handle multi-turn interactive scenarios, thereby enhancing the personalization and effectiveness of web-based task execution.

4.1 Global-Micro User Profile Construction

To address the challenge that a single profile cannot cover users' diverse interests, we propose a global-micro profile mechanism. This design aims to capture both users' long-term stable preferences and context-specific fine-grained needs. These profiles are constructed offline, and can be efficiently retrieved and injected as needed before the user interacts with the web agent.

Global Profile The Global Profile is designed to capture users' long-term stable and task-independent general attributes. To achieve this, we process the user's historical behaviour H using an LLM, following (Cai et al. 2024). The LLM is guided to act as a profile analysis expert and to extract following key personalized dimensions: 1) Basic Information, including inferred demographic details such as gender, age, and occupation; 2) Shopping Preferences, covering price sensitivity, shopping interests, and brand preferences; and 3) Behavioral Tendencies, describing interaction patterns, such as interaction complexity, tone, and focus areas. Through a further decoupling guideline in prompt, this information is separated into positive preferences ρ^{+}_{global}, which represent general interests, and negative preferences ρ^{-}_{global}, which define areas to avoid in all scenarios. Together, these attributes and preferences form the global preference p_{global}.

Fine-grained Micro-Profile Library To overcome the limitations of existing methods in capturing diverse user interests, we further disentangle the user's long behavioral sequence to identify their diversified interests.

First, we perform category-aware interest clustering. We adopt an adaptive bucketing strategy to split the behavioral sequence into subsets according to the categories of historical records, resulting in a set of interest clusters $C = \{c_1, c_2, \ldots, c_N\}$, with each represented by a center vector $\bar{\mathbf{e}}_i$ (the average of its item embeddings). For users with too few interactions ($|H| \leq \alpha$), we skip bucketing and only retain the global profile. This strategy reduces topic fragmentation while preserving the integrity of cross-category compound interests.

We then perform micro-profile extraction. For each interest cluster c_i, we randomly sample up to l_c representative interactions to serve as context, and instruct the LLM to generate a corresponding micro-profile $p_{micro,i}$. As with the global profile procedure, the prompt includes a decoupling instruction that requires the output to be divided into "domain recommendations" and "domain avoidances", corresponding to positive and negative preferences. Thus, each micro-profile $p_{micro,i}$ is decomposed into a positive preference component $\rho^{+}_{micro,i}$ (preferences in the specific domain) and a negative preference component $\rho^{-}_{micro,i}$ (domain-specific materials or functionalities to avoid). Ultimately, these constitute a retrievable micro-profile library $\mathcal{P}micro = \{(\rho^{+}_{micro,1}, \rho^{-}_{micro,1}), \ldots, (\rho^{+}_{micro,N}, \rho^{-}_{micro,N})\}$. This design enables more fine-grained and robust user profiling, enhancing the adaptability of downstream personalized web tasks. This micro-profile library, together with the global

profile $pglobal$, constitutes the complete user profile P_u.

Context-aware Profile Retrieval To overcome the limited expressiveness of single profiles and the inefficiency of global memory retrieval, we introduce a context-aware profile retrieval mechanism.

Driven by the user's goal instruction, this mechanism retrieves the most relevant fine-grained preferences, enabling the agent's initial decisions to incorporate both long-term and task-specific considerations. At the start of each session ($t = 0$), the agent encodes the user's primary goal instruction I as a query embedding $\mathbf{e}_q$ and then performs efficient approximate nearest neighbor (ANN) search (Aumüller, Bernhardsson, and Faithfull 2020) over the pre-stored N interest cluster center vectors $\{\bar{\mathbf{e}}_1, \ldots, \bar{\mathbf{e}}_N\}$:

$$c^* = \underset{c_i \in \{c_1,\ldots,c_N\}}{\operatorname{argmax}} \ \operatorname{sim}(\mathbf{e}_q, \bar{\mathbf{e}}_i),$$

thereby identifying the interest cluster c^* most relevant to the current task context and its corresponding positive micro-profile $\rho^+_{micro,}$. Next, the global negative preference ρ^-_{global} is concatenated with the selected positive micro-preference $\rho^+_{micro,}$ to form the initial context profile $P_{ctx,0} = \operatorname{concat}(\rho^-_{global}, \rho^+_{micro,})$. Here, ρ^-_{global} functions as a universal "safety filter," while $\rho^+_{micro,*}$ provides local positive guidance for the current task. This "global negative filtering + local positive guidance" design allows $P_{ctx,0}$ to maintain consistent long-term avoidance zones while injecting task-relevant fine-grained preferences, thus significantly improving agent decision accuracy. Additionally, this mechanism reduces irrelevant retrieval and system latency.

4.2 Adaptive Agent Decision-Making

To handle evolving user intent and the accumulation of early misunderstandings in multi-turn interactions, we propose an adaptive agent decision-making framework based on multi-source contextual inputs, as illustrated in Figure 3. At each step t, the agent's policy $\pi(a_t|S_t)$ conditions on a belief state representation S_t and selects an action from the space $\mathcal{A}$. Here, S_t serves as the agent's internal belief state, summarizing latent user intent that cannot be directly observed (in contrast, o_t denotes the observable feedback from the environment), allowing the policy to make informed decisions under partial observability. S_t is inspired by cognitive science and constructed by concatenating the user's instruction I, the real-time contextual profile $P_{ctx,t}$, and the session history $C_{s,t}$, enabling the agent to dynamically track the user's preferences.

Adaptive Contextual User Profile To address the dynamic nature of user interests and intents evolving during interactions, we design an adaptive contextual profile $P_{ctx,t}$.

Specifically, the contextual user profile $P_{ctx,t}$ takes the previous profile $P_{ctx,t-1}$ as a prior and combines it with the latest interaction pair (a_{t-1}, o_t). Through an LLM-driven Dynamic Profile Update (DPU) module $\mathcal{G}_{\text{update}}$, the profile is dynamically edited, supplemented, and corrected, enabling adaptive evolution:

$$P_{ctx,t} = \mathcal{G}_{\text{update}}(P_{ctx,t-1}, (a_{t-1}, o_t)).$$

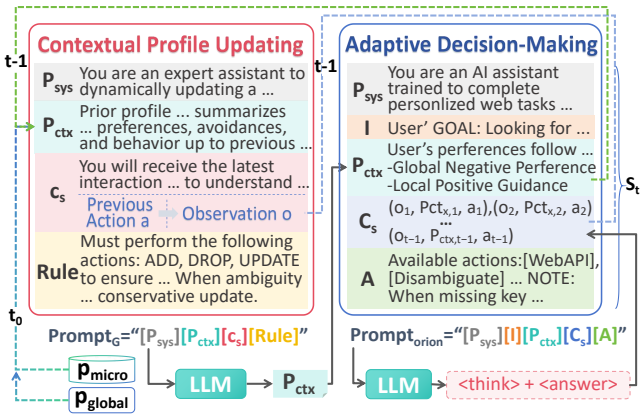

Figure 3: Illustration of the adaptive decision-making framework and user profile updating.

Meanwhile, $P_{ctx,t}$ is appended to the session history $C_{s,t}$. This design effectively improves the coverage and timeliness of user profile information, allowing the agent to dynamically and accurately model and track user preferences.

Enhanced User-Aware Action Space To handle cumulative misunderstandings arising from ambiguous initial user instructions, Orion expands the action space $\mathcal{A}$ to include: 1) predefined web API function calls, enabling concrete task execution, and 2) Proactive Disambiguation (PD) actions, allowing the agent to proactively seek clarification when ambiguity or missing information is detected. This mechanism leverages information-gain-driven interactions to dynamically eliminate decision blind spots and enhance personalization.

With this contextual input and expanded action space, the agent policy π dynamically selects the optimal action, effectively addressing core challenges in complex web interactions and enabling personalized web experiences.

4.3 End-to-End Policy Learning

To enable Orion to efficiently learn to execute tasks through multi-turn interactions, we propose a two-stage end-to-end learning framework. First, SFT establishes foundational reasoning and action patterns. Then, online RL in dynamic environments further aligns and optimizes the policy, thereby enhancing both generalization and personalized adaptability.

Behavior Cloning via SFT To handle cold start and the scarcity of multi-turn personalized web agent data, we propose a bootstrap data generation method via role-playing simulation. Specifically, first, user profiles P_u and corresponding real goals are extracted from the Amazon Review dataset, and a strong LLM teacher model is used to generate natural, ambiguous initial user instructions I via inverse prompting. Next, two information-isolated simulators are designed for automated interaction: (1) The user simulator has access to the contextual user profile $P_{ctx,0}$ and the final answer. According to a controlled disclosure strategy, key information is only released step by step when re-

quested by the agent, avoiding information leakage; (2) The agent simulator can only access the belief state S_t, the goal I, and available tools $\mathcal{A}$, and at each step first explicitly outputs a reason (<think> tag), then outputs the action a_t. The two simulators interact alternately until the agent determines the task is completed or the maximum number of steps is reached, producing structured multi-turn reasoning data:

$$\langle I, \{S_t \mapsto (\text{think}_t, a_t)\}_{t=0}^{T-1} \rangle$$

With this bootstrap corpus of 400 trajectories, the model is fine-tuned under supervision using standard autoregressive loss, enabling it to acquire a "think-then-act" structured decision paradigm and resulting in a base policy π_{SFT} for initializing subsequent RL training.

Online Reinforcement Learning To enhance adaptability in real interactive scenarios, we train Orion via online RL, where the agent learns directly from its own interactive experience rather than a static dataset, a process detailed in Algorithm 1.

Algorithm 1: Online RL for Orion

Require: Initial policy π_{SFT}
for each epoch **do**
  $\mathcal{D} \leftarrow \emptyset$
  for each rollout **do**
    Initialize state S.
    while task not completed **do**
      $a \sim \pi(S)$; $o, R \leftarrow$ Simulator(a).
      $S \leftarrow$ UpdateState(S, a, o). {Updates P_{ctx} part of S}
      Add (S, a, R, o) to $\mathcal{D}$.
    end while
  end for
  Normalize returns G within grouped $\mathcal{D}$.
  Optimize π.
end for
Return optimized policy π

(1) Interactive Trajectory Rollout The agent actively probes the environment to gather a batch of trajectories $\mathcal{D}$. At each step, based on the current belief state S_t, the agent generates a structured action a_t, executes it, and receives observation o_t and reward R_t, dynamically updating the belief state and forming a perception-action loop.

A key challenge in guiding the learning process is credit assignment: in long interactions, a sparse final reward offers no clue as to which specific action led to failure. To provide step-wise guidance, we design a hybrid reward function:

$$R_t = R_{\text{task}} + R_{\text{alignment}}$$

- **Task Success Reward (R_{task}):** A sparse binary reward issued at the end of the interaction. The agent receives +1 if and only if it successfully achieves the user's final goal; otherwise, the reward is 0.
- **Personalized Alignment Reward ($R_{\text{alignment}}$):** A dense reward assessed at each step, quantifying the alignment between action a_t and its current belief state S_t. We use an independent "LLM-as-a-Judge" model to compute this reward by comparing action details against user preferences, outputting a normalized score in $[-1, 1]$: +1 indicates perfect alignment, with the action precisely meeting the user's core preferences; -1 indicates severe violation, where the action touches upon the user's negative preferences or "taboos"; 0 indicates a neutral or unrelated action. This signal directly guides the agent to learn how to leverage the user profile for reasoning. While this provides a scalable reward, we acknowledge the LLM judge may have inherent biases, and its scores are an approximation of true user preferences.

(2) Policy Optimization The experiential data collected in D is informative, but the hybrid reward structure introduces high variance in the returns. To address this, we employ GRPO algorithm (Guo et al. 2025), initializing from π_{SFT}. By normalizing rewards within each group of trajectories, it stabilizes the policy updates and enhances sample efficiency, ultimately yielding a robust personalized policy π.

5 Experiments

5.1 Experimental Setup

Dataset To systematically evaluate Orion's capability in complex multi-turn personalized web agent scenarios, we extend and reconstruct the PersonalWAB (Cai et al. 2024) benchmark based on the Amazon Review dataset, building a new experimental dataset. The user instructions in this dataset are designed to be closer to real-world scenarios, mostly open-ended, ambiguous, or incomplete, thus providing a more rigorous test of the agent's ability to clarify user intent, proactively interact, and dynamically model personalization. The dataset covers 1,000 real users, each with a rich interaction history (on average, about 30 records per user), fully reflecting users' long-term and diverse personalized needs.

The experiments span three personalized web task scenarios: 1)Product Search: The agent incrementally infers the user's true intent from vague or incomplete requirements via multi-turn interaction, incorporating user history for intelligent ranking and recommendation during product filtering; 2) Personalized Recommendation: The agent combines user history with the current context to track interest drift and context changes in real time, proactively guides users to explore new interests, and dynamically adjusts recommendation strategies based on feedback, achieving both global and context-specific personalization. 3)Review Generation: The agent models the user's linguistic style, sentiment, and focus in historical reviews to generate highly personalized product reviews, and can iteratively refine text content through multi-turn feedback.

Metrics To comprehensively evaluate multi-turn personalized interaction, we adopt three metrics:

- **Function Accuracy (F. Acc.):** Measures whether the agent selects the correct web function and generates valid

Method	Product Search			Personalized Recommendation			Review Generation		
	F.Acc(%)	R.Acc(%)	Steps	F.Acc(%)	R.Acc(%)	Steps	F.Acc(%)	R.Acc(%)	Steps
GPT-4.1 (Raw)	61.5	20.2	7.5	31.3	1.1	8.6	54.5	19.3	4.6
GPT-4.1 (CoT)	71.2	35.3	6.9	43.5	1.4	7.4	68.7	27.5	4.0
GPT-4.1 (ReAct)	74.8	36.5	6.8	44.6	2.1	7.3	70.2	28.4	3.9
GPT-4.1 (ChatShop)	78.5	38.0	7.2	47.5	2.3	7.6	72.5	29.0	4.1
GPT-4.1 (PUMA)	80.5	42.8	6.5	50.2	2.6	7.1	75.5	30.5	3.6
DeepSeek-V3 (Raw)	62.1	19.2	7.8	29.5	0.9	9.0	51.2	14.2	5.8
DeepSeek-V3 (CoT)	71.2	35.3	7.2	41.5	1.5	7.7	66.5	26.4	4.2
DeepSeek-V3 (ReAct)	73.5	36.4	7.0	43.7	1.6	7.6	68.4	27.3	4.0
DeepSeek-V3 (ChatShop)	76.5	37.5	7.4	45.8	1.9	7.9	70.5	28.1	4.2
DeepSeek-V3 (PUMA)	78.5	40.5	6.9	48.5	2.2	7.4	73.5	29.4	3.9
Qwen2.5 (Raw)	61.4	19.7	7.9	30.9	1.0	8.6	51.4	17.2	5.1
Qwen2.5 (CoT)	65.3	30.4	7.8	35.4	1.1	8.2	60.4	22.2	4.3
Qwen2.5 (ReAct)	67.1	31.5	7.7	37.8	1.2	8.1	62.5	23.3	4.2
Qwen2.5 (ChatShop)	70.5	32.8	8.0	40.2	1.5	8.4	64.8	24.0	4.4
Qwen2.5 (PUMA)	72.5	35.3	7.4	42.5	1.8	7.9	67.5	25.4	4.0
Qwen2.5-7B-Instruct (SFT)	90.5	55.5	5.1	70.5	8.2	5.6	90.5	40.5	2.6
Qwen2.5-7B-Instruct (SFT + PPO)	92.5	58.5	4.8	75.5	10.3	5.3	93.5	45.5	2.4
Orion (Ours)	**94.5**	**62.5**	**4.5**	**80.5**	**19.8**	**4.8**	**95.5**	**50.5**	**2.2**

Table 1: Main results on the three core tasks. We report Functional Accuracy (F. Acc.) and Result Accuracy (Res. Acc.) as percentages (%), and Average Steps.

parameters at each step, indicating instruction comprehension and tool invocation ability.

- **Result Accuracy (Res. Acc.):** Evaluates the personalized quality of the final output. For search and recommendation, this is based on the ranking of the target item in the results; for content generation, it is measured by the semantic similarity between the generated and reference texts.
- **Average Steps:** Records the total number of interaction steps required to complete the task, including API calls and clarification dialogs. Fewer steps indicate higher efficiency.

Compared Methods and Implementation Details To comprehensively evaluate our Orion, we conduct comparisons with current mainstream prompt-based and fine-tuning-based methods. For the prompt-based category, we select the latest state-of-the-art (SOTA) LLMs as backbones: GPT-4.1 (Achiam et al. 2023), Qwen2.5 (72B) (Qwen et al. 2025), and DeepSeek-V3 (Liu et al. 2024). For each model, we apply five different prompting strategies: Raw (zero-shot) prompting, CoT (Chain-of-Thought) prompting (Wei et al. 2022), ReAct prompting (Yao et al. 2023), ChatShop method (Chen, Wiseman, and Dhingra 2024) (as a representative of non-personalized web agents), and the Personalized Prompting (Cai et al. 2024) strategy based on the PUMA framework, representing the current SOTA of prompt engineering–based personalized solutions.

In Global-Micro Profile construction, all profiles are constructed using GPT-4o-mini. The clustering threshold α and context size l_c are set to 15 and 6, respectively. For fine-tuning methods, all experiments use the Qwen2.5-7B-Instruct model. Our baselines are SFT and SFT + PPO (Schulman et al. 2017), which applies standard Proximal Policy Optimization to the SFT model. All RL experiments are conducted using the veRL framework (Sheng et al. 2025). For SFT, we use the AdamW optimizer with a learning rate of 2e-6, batch size of 1, and train for 3 epochs. For online RL (PPO and GRPO), the generation temperature is set to 1.0 for the review generation task and 0.6 for other tasks, the clip range is 0.2, and reward scaling is 1.0. All experiments are conducted on 2 NVIDIA H100 80G GPUs. Both RL methods (PPO and GRPO) are initialized from the same SFT model and share the same reward function for fair comparison. The personalized alignment reward is calculated by GPT-4o-mini acting as the "LLM-as-a-Judge".

5.2 Main Results

Table 1 summarizes the experimental results for Orion and all baselines. Orion achieves SOTA performance across all key metrics, significantly outperforming both prompt-based and fine-tuning approaches.

Notably, in the most challenging personalized recommendation task, Orion reaches a result accuracy (R. Acc) of 19.8%, nearly 8 times higher than the best prompt-based baseline—PUMA with GPT-4.1 (2.6%). This significant improvement is due to our global-micro profile design and context-aware retrieval mechanisms, which allow the agent to focus on fine-grained user preferences, rather than relying on static, averaged profiles as in prompt-based methods. Traditional prompt engineering struggles with vague and open-ended instructions, showing clear bottlenecks in both function selection and final result quality. Function accuracy is generally below 80%, result accuracy is even lower, and average interaction steps are higher, all indicating inefficiency in handling complex multi-turn personalized scenarios.

In fine-tuning comparisons, Orion also demonstrates clear advantages. For personalized recommendation, Orion boosts result accuracy from 8.2% (SFT) to 19.8%, nearly doubling SFT+PPO (10.3%). Similar trends appear in product search and review generation: product search function accuracy increases from 90.5% (SFT) to 94.5%, and review

generation result accuracy rises from 45.5% (SFT+PPO) to 50.5%. Orion also consistently reduces average interaction steps—for personalized recommendation, steps decrease from 5.6 to 4.8, showing that end-to-end policy learning, proactive disambiguation, and adaptive profile tracking help agents achieve personalized goals more effectively, enhancing user experience.

Compared with traditional PPO fine-tuning, Orion achieves a 19.8% result accuracy in personalized recommendation, a 9.5 percentage point improvement over PPO. We attribute this superior performance to our choice of GRPO. This result demonstrates that GRPO's group normalization is a more stable and effective optimization strategy for our hybrid reward setting, validating our policy learning method.

Overall, these substantial improvements are due to our innovations in global-micro profile design, context-aware personalized retrieval, adaptive tracking and proactive disambiguation, and efficient GRPO-based RL. Together, they enable Orion to set a new performance for personalized web agents in multi-turn and complex scenarios.

5.3 Ablation Study

To verify the effectiveness of each core component in the Orion framework, we conduct a series of ablation experiments, as shown in Table 2.

Method	F. Acc.(%)	Res. Acc.(%)	Avg. Steps(%)
Orion (Full Model)	**90.2**	**44.3**	**3.83**
w/o RL (SFT only)	83.8	34.7	4.4
w/o G-M Profile	79.7	30.8	5.1
w/o DPU	81.4	31.6	4.8
w/o PD	84.9	32.1	6.4

Table 2: Ablation study of the Orion's components. We report the average performance across all three tasks.

Removing the RL module and retaining only the SFT model (w/o RL) leads to drops of 6.4% and 9.6% in function accuracy (F. Acc.) and result accuracy (R. Acc.), respectively. This indicates that while SFT injects basic paradigms through behavior cloning, online RL further benefits the agent in learning and optimizing personalized policies in dynamic environments.

Removing the hierarchical profile structure (w/o G-M Profile) and using a single profile results in the largest drop in result accuracy, decreasing by 13.5% compared to the full model. This shows that a single, averaged profile cannot capture users' fine-grained interests, whereas our global-micro profile design and context-aware retrieval method effectively enhance context-aware personalization.

If the profile is not updated during multi-turn interactions (w/o DPU), result accuracy drops by 12.7 percentage points relative to the full model—a significant performance loss. This validates the core challenge discussed in the introduction, namely that user intent evolves dynamically throughout interaction. The DPU mechanism ensures that the agent can continuously track changing user preferences by editing and correcting the context profile in real time, thus enabling more accurate decisions.

Restricting the model's ability to proactively ask questions (w/o PD) leads to substantial decreases in function accuracy and result accuracy (by 5.3 and 12.2 percentage points, respectively), and the average number of interaction steps increases markedly from 3.83 to 6.4. This demonstrates that proactive clarification actions, when facing ambiguous user instructions, help prevent the agent from falling into ineffective trial-and-error loops, improving both task success rate and interaction efficiency.

In summary, these results provide further evidence for the effectiveness of Orion's core design and highlight the distinct contribution of each component.

5.4 Parameter Sensitivity Analysis

To determine Orion's key parameters, we conducted a sensitivity analysis.

Clustering Threshold for Micro-Profiles We examine the impact of the clustering threshold α (see Figure 4). If the threshold is set too low, even users with sparse interactions are forced into clusters, which can produce meaningless or noisy clusters and reduce result accuracy. If the threshold is too high, users with moderate interaction histories cannot be finely modeled, limiting performance. The results indicate that $\alpha = 15$ achieves the best balance between filtering sparse users and effective modeling.

Context Size for Micro-Profile Extraction We then investigate the number of interaction samples l_c used for micro-profile extraction (see Figure 4). With fewer samples, context is insufficient to generate high-quality micro profiles, degrading performance. When l_c increases to 6, performance peaks. Further increases provide limited or even negative effects, likely due to the inclusion of redundant or non-representative interactions and increased LLM processing cost. Thus, we choose $l_c = 6$ as the configuration.

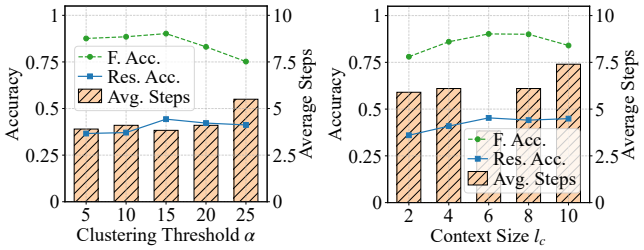

Figure 4: Hyperparameter sensitivity analysis.

6 Conclusion and Future Work

This paper presents Orion, a personalized web agent framework based on global-micro profiling, effectively addressing the core challenges of user interest diversity modeling and dynamic intent tracking in multi-turn web interactions. Leveraging hierarchical profiles, context-aware retrieval, adaptive decision-making, and end-to-end learning, Orion achieves leading performance over three tasks. In the future, we plan to extend Orion to broader web scenarios and multilingual tasks, and to explore privacy protection mechanisms for personalized web agents.

Acknowledgments

This work was supported by Grant No. E4GZ08020403. We also thank the Artificial Intelligence Chip and Systems R&D Center, Institute of Microelectronics of the Chinese Academy of Sciences, for their support.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aumüller, M.; Bernhardsson, E.; and Faithfull, A. 2020. ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems*, 87: 101374.

Cai, H.; Li, Y.; Wang, W.; Zhu, F.; Shen, X.; Li, W.; and Chua, T.-S. 2024. Large Language Models Empowered Personalized Web Agents. *Proceedings of the ACM on Web Conference 2025*.

Chen, S.; Wiseman, S.; and Dhingra, B. 2024. Chatshop: Interactive information seeking with language agents. *arXiv preprint arXiv:2404.09911*.

Deng, X.; Gu, Y.; Zheng, B.; Chen, S.; Stevens, S.; Wang, B.; Sun, H.; and Su, Y. 2023. MIND2WEB: towards a generalist agent for the web. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Foster, D. J.; Block, A.; and Misra, D. 2024. Is behavior cloning all you need? understanding horizon in imitation learning. *Advances in Neural Information Processing Systems*, 37: 120602–120666.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Humphreys, P. C.; Raposo, D.; Pohlen, T.; Thornton, G.; Chhaparia, R.; Muldal, A.; Abramson, J.; Georgiev, P.; Santoro, A.; and Lillicrap, T. 2022. A data-driven approach for learning to control computers. In *International Conference on Machine Learning*, 9466–9482. PMLR.

Kurniawati, H. 2021. Partially Observable Markov Decision Processes (POMDPs) and Robotics. arXiv:2107.07599.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, A.; Wu, Z.; Michael, J.; Suhr, A.; West, P.; Koller, A.; Swayamdipta, S.; Smith, N. A.; and Choi, Y. 2023. We're afraid language models aren't modeling ambiguity. *arXiv preprint arXiv:2304.14399*.

Nguyen, D.; Chen, J.; Wang, Y.; Wu, G.; Park, N.; Hu, Z.; Lyu, H.; Wu, J.; Aponte, R.; Xia, Y.; et al. 2025. GUI Agents: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2025*, 22522–22538. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Ning, L.; Liang, Z.; Jiang, Z.; Qu, H.; Ding, Y.; Fan, W.; Wei, X.-y.; Lin, S.; Liu, H.; Yu, P. S.; and Li, Q. 2025. A Survey of WebAgents: Towards Next-Generation AI Agents for Web Automation with Large Foundation Models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, 6140–6150. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714542.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Sheng, G.; Zhang, C.; Ye, Z.; Wu, X.; Zhang, W.; Zhang, R.; Peng, Y.; Lin, H.; and Wu, C. 2025. HybridFlow: A Flexible and Efficient RLHF Framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, 1279–1297. ACM.

Shi, Y.; Xu, W.; Zhang, Z.; Zi, X.; Wu, Q.; and Xu, M. 2025. Personax: A recommendation agent oriented user modeling framework for long behavior sequence. *arXiv preprint arXiv:2503.02398*.

Stengel-Eskin, E.; Guallar-Blasco, J.; Zhou, Y.; and Van Durme, B. 2023. Why Did the Chicken Cross the Road? Rephrasing and Analyzing Ambiguous Questions in VQA. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10220–10237. Toronto, Canada: Association for Computational Linguistics.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2): 121101.

Yao, S.; Chen, H.; Yang, J.; and Narasimhan, K. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35: 20744–20757.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629.

Zhou, S.; Xu, F. F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.