A Survey of WebAgents: Towards Next-Generation AI Agents for Web Automation with Large Foundation Models

Liangbo Ning*
The Hong Kong Polytechnic University
Hong Kong SAR
BigLemon1123@gmail.com

Ziran Liang*
The Hong Kong Polytechnic University
Hong Kong SAR
ziran.liang@connect.polyu.hk

Zhuohang Jiang
The Hong Kong Polytechnic University
Hong Kong SAR
zhuohang.jiang@outlook.com

Haohao Qu
The Hong Kong Polytechnic University
Hong Kong SAR
haohao.qu@connect.polyu.hk

Yujuan Ding
The Hong Kong Polytechnic University
Hong Kong SAR
dingyujuan385@gmail.com

Wenqi Fan†
The Hong Kong Polytechnic University
Hong Kong SAR
wenqifan03@gmail.com

Xiao-yong Wei
The Hong Kong Polytechnic University
Hong Kong SAR
cs007.wei@polyu.edu.hk

Shanru Lin
City University of Hong Kong
Hong Kong SAR
lllam32316@gmail.com

Hui Liu
Michigan State University
Michigan, USA
liuhui7@msu.edu

Philip S. Yu
University of Illinois at Chicago
Chicago, USA
psyu@uic.edu

Qing Li
The Hong Kong Polytechnic University
Hong Kong SAR
qing-prof.li@polyu.edu.hk

Abstract

With the advancement of web techniques, they have significantly revolutionized various aspects of people's lives. Despite the importance of the web, many tasks performed on it are repetitive and time-consuming, negatively impacting the overall quality of life. To efficiently handle these tedious daily tasks, one of the most promising approaches is to advance autonomous agents to incorporate human-like intelligence based on Artificial Intelligence (AI) techniques, referred to as **AI Agents**. AI Agents offer significant advantages in handling such tasks since they can operate continuously without fatigue or performance degradation. Therefore, leveraging AI Agents – termed **WebAgents** in the context of web – to automatically assist people in handling tedious daily tasks can dramatically enhance productivity and efficiency. Recently, Large Foundation Models (**LFMs**) containing billions of parameters have exhibited human-like language understanding and reasoning capabilities, showing proficiency in performing various complex tasks. This naturally raises the question: '*Can LFMs be utilized to develop powerful AI Agents that automatically handle web tasks, providing significant convenience to users?*' To fully explore the potential of LFMs, extensive research has emerged on WebAgents designed to complete daily web tasks according to user instructions, significantly enhancing the convenience of daily human life. In this survey[1], we comprehensively review existing research studies on WebAgents across three key aspects: architectures, training, and trustworthiness. Additionally, several promising directions for future research are explored to provide deeper insights.

CCS Concepts

• **Computing methodologies → Intelligent agents**; • **Information systems → Web applications**.

Keywords

WebAgents, Large Foundation Models, AI Agents, AI Assistants, Prompting, Pre-training, Fine-tuning.

ACM Reference Format:
Liangbo Ning, Ziran Liang, Zhuohang Jiang, Haohao Qu, Yujuan Ding, Wenqi Fan, Xiao-yong Wei, Shanru Lin, Hui Liu, Philip S. Yu, and Qing Li. 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), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3711896.3736555

1 Introduction

As the web has rapidly evolved, it has profoundly transformed various aspects of people's lives, including information access [21, 47, 50], shopping experiences [9, 43], and communications [10, 82]. For instance, the web serves as the largest knowledge repository to date, offering instant access to news [41, 124], academic papers (e.g., ArXiv [17]), and encyclopedias (e.g., Wikipedia [96, 105]), enabling individuals to freely acquire desired information. This advancement has eliminated geographical barriers, providing people in remote areas with access to critical resources in education, healthcare, and law. Despite the importance of the web, many daily tasks we perform on it are repetitive and extremely time-consuming.

*Both authors contributed equally to this research.
†Corresponding author: Wenqi Fan, Department of Computing, and Department of Management and Marketing, The Hong Kong Polytechnic University.
[1] The long version of this survey can be found at: *https://arxiv.org/abs/2503.23350*.

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3736555

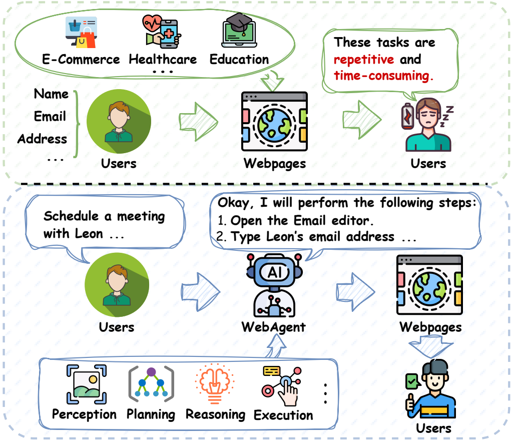

E-Commerce
Healthcare
Education
...
These tasks are repetitive and time-consuming.
Name
Email
Address
...
Users
Webpages
Users
Schedule a meeting with Leon ...
Okay, I will perform the following steps:
1. Open the Email editor.
2. Type Leon's email address ...
AI
Users
WebAgent
Webpages
Perception
Planning
Reasoning
Execution
Users

Figure 1: Illustration of basic web tasks and the pipeline of WebAgents. Given the user instruction, WebAgents autonomously complete tasks by perceiving the environment, reasoning action sequences, and executing interactions.

For example, as shown in Figure 1, when registering accounts on various platforms or filling out different application forms, we are often required to repeatedly enter the same personal information, such as our name, contact details, and address. Similarly, when purchasing a product, we need to compare numerous options, review their ratings and prices, and ultimately decide on the final purchase. To effectively execute tedious daily tasks, one of the most promising techniques is to develop automatic agents embedded with human intelligence by taking advantage of Artificial Intelligence (AI) techniques, known as **AI Agents**. In addition, AI Agents can execute tasks continuously without fatigue or performance degradation [39], ensuring reliability in repetitive workflows. Therefore, leveraging AI Agents – *termed **WebAgents** in the context of web* – to assist people in handling tedious daily tasks automatically can extremely enhance productivity and efficiency, thereby further improving their quality of life.

Recently, large foundation models (**LFMs**) with billions of parameters, trained on massive data, have exhibited emergent human-like capabilities such as comprehension and reasoning, revolutionizing various domains including healthcare [21, 29, 79], e-commerce [23, 74], and AI4Science [22, 33]. For example, LFMs are integrated with protein data to capture the foundational protein knowledge, enabling better understanding and generation of protein structures, which can significantly advance the development of drug discovery and disease mechanism research [22]. The human-like reasoning capabilities of LFMs are also leveraged in recommender systems (RecSys) to provide better item recommendations, significantly enhancing user online experience [107, 128]. By leveraging their extensive open-world knowledge, advanced instruction-following, and language comprehension and reasoning abilities, LFMs exhibit proficiency in simulating human-like behaviors to execute a variety of complex tasks. This naturally raises the promising topic: '*Can LFMs be utilized to develop powerful AI Agents that automatically handle web tasks, providing significant convenience to users?*'

To fully explore the potential of LFMs, recent efforts have been made to advance LFM-empowered **WebAgents** to complete various web tasks according to user instructions [39]. For instance, the recent debut of a novel AI Agent framework named *AutoGPT* has attracted significant interest from both academic and industrial communities, which exhibits impressive capabilities in autonomously handling complex tasks across both work and daily environments [87]. Unlike chatbots, AutoGPT can plan and execute complex tasks independently, performing automated searches and multi-step actions without requiring ongoing user instructions and supervision. In this context, as illustrated in Figure 1, users only need to provide a natural language instruction, such as '*Schedule a meeting with Leon at Starbucks on November 23, 2024, at 4:00 pm via email.*' WebAgents can autonomously open the 'Email' application, retrieve Leon's email address, compose the email, and send it, thereby automating the entire scheduling process and greatly enhancing the convenience of daily life. Given the remarkable progress in developing LFM-empowered WebAgents and the growing number of related studies, there is a pressing need for a systematic review of recent advances in this field.

To bridge this gap, this survey provides a comprehensive overview of WebAgents by summarizing representative methods from the perspectives of architecture, training, and trustworthiness. Specifically, in Section 2, we review existing studies based on the three processes of WebAgents: perception, planning & reasoning, and execution. Next, we summarize two crucial aspects (i.e., data and training strategies) in the training of WebAgents in Section 3. After that, we review studies that focus on investigating the trustworthy WebAgents, including their safety & robustness, privacy, and generalizability, in Section 4. Finally, in Section 5, we discuss promising future research directions in WebAgents.

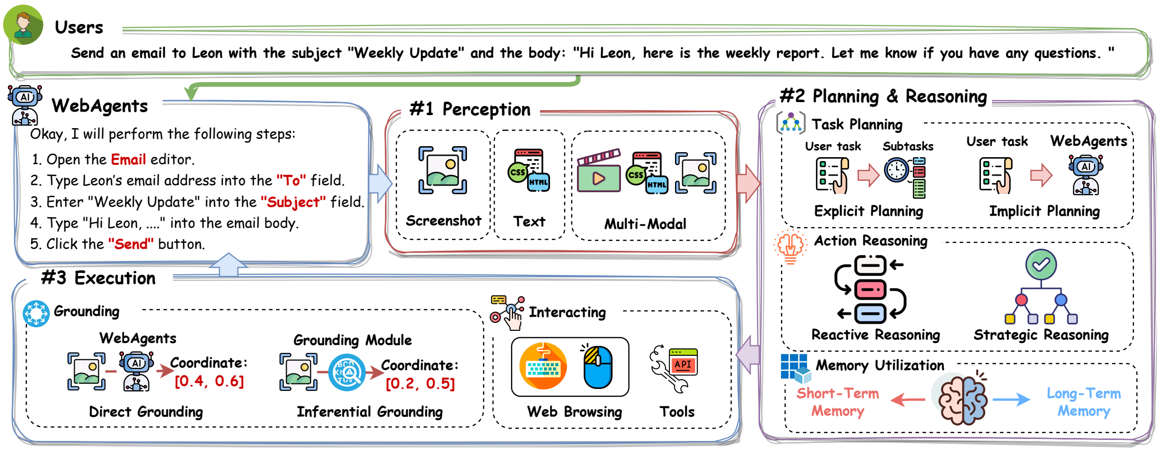
2 WebAgent Architectures

There are three crucial and consecutive processes for WebAgents to fulfill user commands: **1) Perception** requires WebAgents to accurately observe the current environment, **2) Planning & Reasoning** require WebAgents to analyze the current environment, interpret user-given tasks, and predict reasonable next actions, and **3) Execution** requires that WebAgents perform the generated actions and interact with the environment effectively. In the following section, we will comprehensively review the important techniques employed by WebAgents during these processes.

2.1 Perception

Typical LFMs merely need to accept user instructions and generate corresponding responses through reasoning. However, WebAgents, operating within complex web environments, are further expected to accurately perceive the external environment and perform behavioral reasoning based on the dynamic environment combined with the user's task. As shown in Figure 2, according to the data modality provided by the environment to WebAgents, we can categorize existing studies into three classes: 1) **Text-based**, 2) **Screenshot-based**, and 3) **Multi-modal** WebAgents.

2.1.1 Text-based WebAgents. With the advancement of large language models (LLMs), extensive studies have been proposed to leverage its human-like understanding and reasoning abilities to



Users

Send an email to Leon with the subject "Weekly Update" and the body: "Hi Leon, here is the weekly report. Let me know if you have any questions. "

WebAgents

Okay, I will perform the following steps:

1. Open the Email editor.
2. Type Leon's email address into the "To" field.
3. Enter "Weekly Update" into the "Subject" field.
4. Type "Hi Leon, ..." into the email body.
5. Click the "Send" button.

Figure 2: Illustration of the overall framework of WebAgents, which contains three crucial processes: *Perception*, *Planning & Reasoning*, and *Execution*. Given the user's command, WebAgents first observe the environmental information during the perception process. Based on the observation, the action is generated in the planning & reasoning process. Finally, WebAgents execute the generated action to complete the user's task.

assist users in addressing complex tasks. Since LLMs can only handle natural language, these WebAgents usually leverage the textual metadata of webpages (e.g., **HTML** and **accessibility trees**) to perceive the environment [45, 67, 69, 129]. For example, MindAct [18] introduces a two-stage framework that combines a fine-tuned small language model (LM) with an LLM to efficiently process large HTML documents, significantly reducing the input size while preserving essential information. This approach enables accurate prediction of both the target element and the corresponding action, effectively balancing efficiency and performance in web-based tasks. Gur et al. [35] introduce an LLM-driven agent that learns from self-experience to complete tasks on real-world webpages. It summarises long HTML documents into task-relevant snippets to extract the environmental information and decomposes user instructions into sub-tasks for effective planning.

2.1.2 Screenshot-based WebAgents. Despite the remarkable success of text-based WebAgents, leveraging the textual metadata of the environment usually fails to align closely with human cognitive processes since the Graphical User Interfaces (GUI) are inherently visual [84, 116]. Additionally, textual representations usually vary across different environments and are verbose, leading to poor generalization abilities and increased computational overhead [127]. Recently, breakthroughs in large vision-language models (VLMs) have significantly enhanced the capabilities of AI systems in processing complex visual interfaces. To leverage the visual understanding capabilities of VLMs, numerous studies have integrated them into WebAgents, utilizing screenshots to perceive the environment [27, 31, 44, 127]. For example, SeeClick [15] only relies on screenshots as observations to predict the next action and enhances the agent's ability to locate relevant visual elements within screenshots by introducing a grounding pre-training process. OmniParser [68] introduces an effective method to parse user interface screenshots into structured elements and enhances GPT-4V's [119] ability to accurately ground actions to specific regions on the screen.

2.1.3 Multi-modal WebAgents. In addition to solely utilizing textual metadata or screenshots to comprehend the environment, numerous studies also leverage multi-modal data, combining their complementary strengths to provide WebAgents with a more comprehensive environmental perception [42, 52, 99]. For instance, MMAC-Copilot [91] integrates GPT-4V for interpreting visual information from screenshots while leveraging Gemini Vision [59] to process and analyze video content, significantly enhancing the model's capabilities in handling multi-modal data. WebVoyager [37] is a multi-modal WebAgent that autonomously completes web tasks end-to-end by processing both screenshots and textual content from interactive web elements. It leverages Set-of-Mark Prompting [117] to overlay bounding boxes of the interactive elements on the webpages, significantly enhancing the agent's decision-making ability and enabling accurate action prediction and execution.

2.2 Planning & Reasoning

Subsequent to the perception of environmental information, WebAgents are generally tasked with determining the appropriate action to execute the user's command. This involves analyzing the current state of the environment and utilizing the reasoning capabilities of LFMs. As shown in Figure 2, there are three subtasks involved in this process: 1) **Task Planning**, which focuses on reorganizing the user's instruction and setting sub-objectives to help WebAgents effectively handle complex user queries; 2) **Action Reasoning**, which guides WebAgents to generate appropriate actions to fulfill the user's commands; and 3) **Memory Utilization**, which equips WebAgents with internal information (e.g., previous actions) or external information (e.g., open-world knowledge from web search) to predict more appropriate actions.

2.2.1 Task Planning. In the context of WebAgents, the objective of task planning is to determine a sequence of steps that the agent should take to complete the user-defined task efficiently and effectively [112]. Based on whether WebAgents explicitly involve