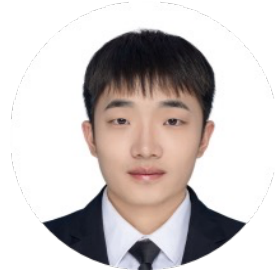


# Towards Autonomous Web Interaction: Empowering WebAgents by LFM



Yujuan Ding



Liangbo Ning



Ziran Liang



Haohao Qu



Wenqi Fan



The Hong Kong Polytechnic University

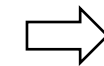
June 9, 2026



# Tutorial Outline

- ⊙ **Part 1: Introduction of WebAgents**
- **Part 2: Preliminaries** of AI Agents and LFM-based WebAgents
- **Part 3: Architectures** of WebAgents
- **Coffee Break**
- **Part 4: Training** of WebAgents
- **Part 5: Trustworthy WebAgents**
- **Part 6: Future directions** of WebAgents

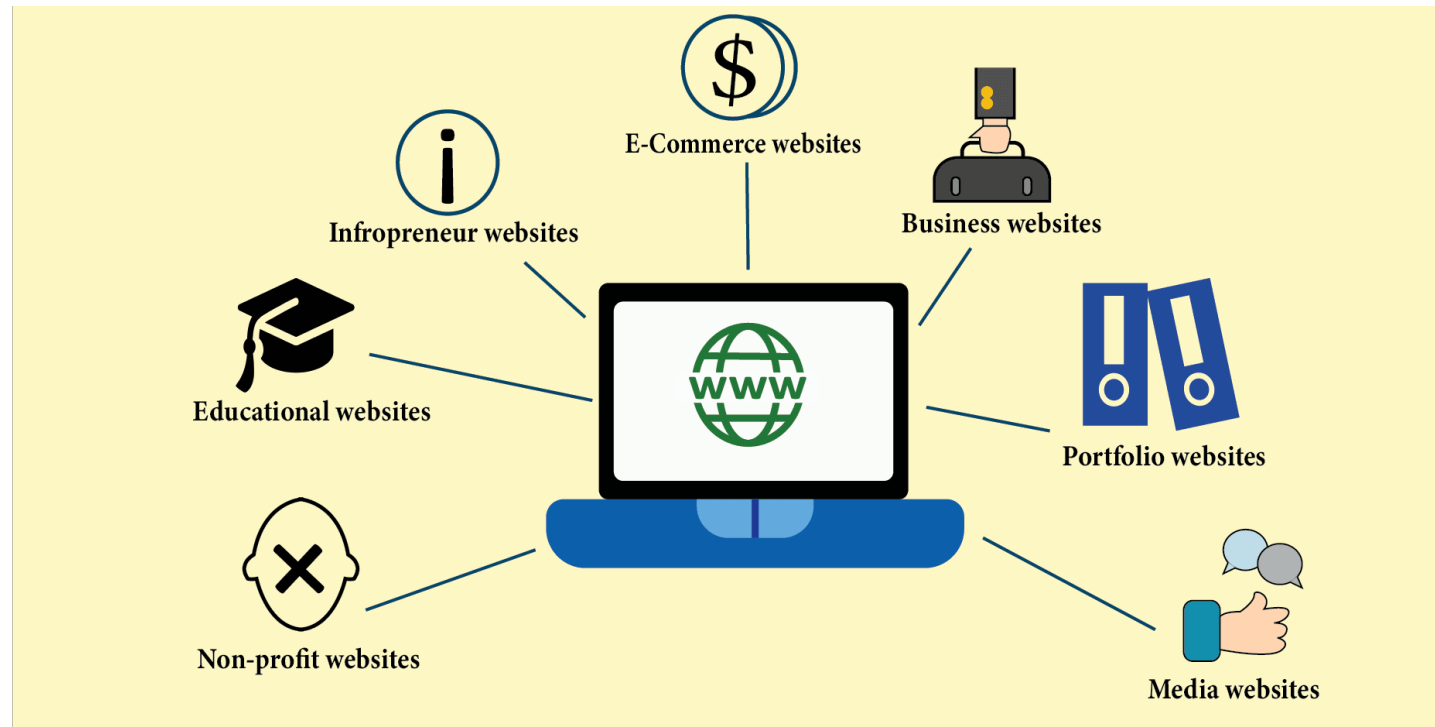
Website of this tutorial  
Check out the slides and more information!



# Web

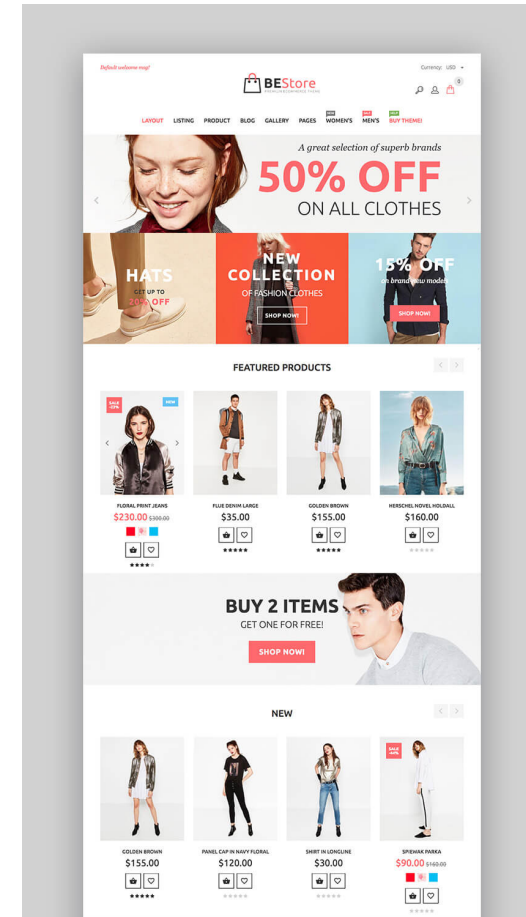
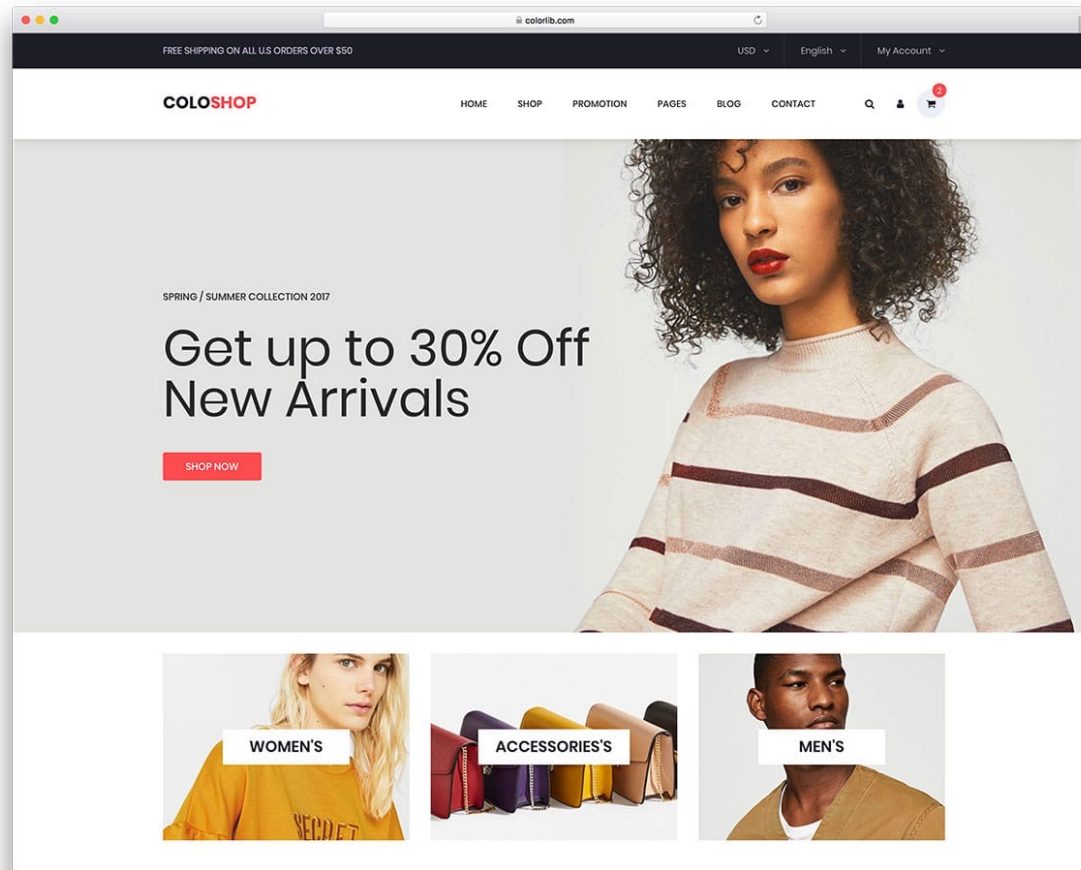
*“The Web does not just connect machines,  
it connects people.”*

— Tim Berners-Lee,  
Inventor of the World Wide Web,  
ACM A.M. Turing Award Laureate 2016.



# Web for E-Commerce

*Web provides a digital platform that enables businesses and consumers to buy and sell goods and services online through websites and applications.*



# Web for Education

*Web facilitates access to learning resources, online courses, and interactive tools for teaching and knowledge sharing.*

The screenshot shows the top navigation bar of the HKU Teaching & Learning website. It includes a search bar and links to various resources like 'Generative AI', 'e-learning Blog', and 'HKU Online Learning & MOOCs'. Below the navigation, there is a breadcrumb trail: '> T&L@HKU > 4-Year UG Curriculum > Common Core > Academic Advising > Experiential Learning > Horizons > CICs > TALIC > Staff'. The main heading reads 'HKU Online Learning MOOCs and Professional Certificate Programs'.

Get access to more than 40 Massive Open Online Courses (MOOCs) and 5 Professional Certificate programs on edX and Coursera, offered by the University of Hong Kong. Our innovative MOOCs bring together top academics, industry leaders, and practitioners from around the world to provide a global learning experience.

Our MOOCs are also integrated into on-campus courses, leveraging the global learning opportunities to enrich students' learning experiences through MOOC-based blended learning modules, flipped classrooms, and SPOCs (small private online courses)."

Learn at your own pace, anytime, anywhere, with bite-sized lessons and interactive activities. Our courses help you bridge knowledge gaps, upskill in specialized areas, and explore new interests. Join our courses on **edX** and **Coursera** now and discover a world of learning opportunities!

- [e-learning Blog](#)
- [How to join HKUx](#)
- [Our Facebook](#)

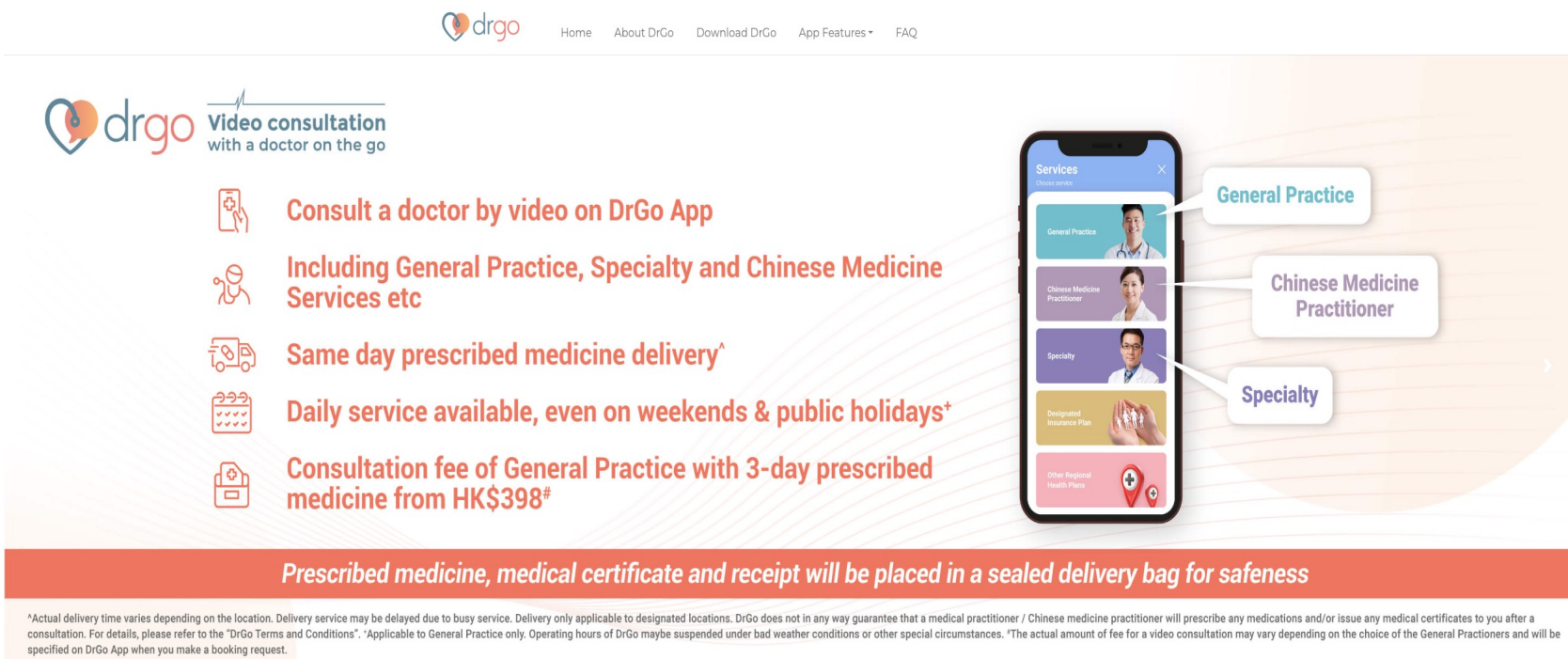
## Professional Certificate Programs

Three professional certificate program cards are displayed. Each card features a header image, a 'Professional Certificate' label, and the program name with the number of courses. The first card is for 'FinTech' (3 Courses), the second for 'FinTech Technologies I' (3 Courses), and the third for 'Advanced Cardiac Imaging' (2 Courses).

The screenshot shows the Wikipedia homepage in Chinese. The main heading is 'WIKIPEDIA 自由的百科全书'. Below it, there are statistics for various languages: 中文 (1,482,000+ 条目), English (7,009,000+ articles), 日本語 (1,462,000+ 記事), Deutsch (3.024.000+ Artikel), Français (2.690.000+ articles), and Português (1.148.000+ artigos). There is also a section for فارسی (1,043,000+ مقاله). A central globe icon is surrounded by these language statistics. At the bottom, there is a search bar with a dropdown menu set to 'ZH' and a search icon.

# Web for Healthcare

*Web supports medical information sharing, telemedicine, patient management, and access to health-related resources and services.*



The screenshot shows the DrGo website with a navigation bar at the top containing the logo and links for Home, About DrGo, Download DrGo, App Features, and FAQ. The main content area features the DrGo logo and the text "Video consultation with a doctor on the go". Below this, there are five key service points, each with an icon: consulting a doctor by video on the DrGo App, including General Practice, Specialty, and Chinese Medicine Services, same-day prescribed medicine delivery, daily service availability on weekends and public holidays, and a consultation fee for General Practice with 3-day prescribed medicine. A central image of a smartphone displays the DrGo app interface with callouts for General Practice, Chinese Medicine Practitioner, and Specialty services. A red banner at the bottom states that prescribed medicine, medical certificates, and receipts will be placed in a sealed delivery bag for safety.

drgo Video consultation with a doctor on the go

- Consult a doctor by video on DrGo App
- Including General Practice, Specialty and Chinese Medicine Services etc
- Same day prescribed medicine delivery\*
- Daily service available, even on weekends & public holidays\*
- Consultation fee of General Practice with 3-day prescribed medicine from HK\$398#

Services

- General Practice
- Chinese Medicine Practitioner
- Specialty
- Designated Insurance Plan
- Other Regional Health Plans

Prescribed medicine, medical certificate and receipt will be placed in a sealed delivery bag for safeness

\*Actual delivery time varies depending on the location. Delivery service may be delayed due to busy service. Delivery only applicable to designated locations. DrGo does not in any way guarantee that a medical practitioner / Chinese medicine practitioner will prescribe any medications and/or issue any medical certificates to you after a consultation. For details, please refer to the "DrGo Terms and Conditions". #Applicable to General Practice only. Operating hours of DrGo maybe suspended under bad weather conditions or other special circumstances. \*The actual amount of fee for a video consultation may vary depending on the choice of the General Practitioners and will be specified on DrGo App when you make a booking request.



The advertisement features a green background with the DrGo logo and the text "Online Doctor Consult eDoctor Anytime Anywhere". A cartoon doctor character is shown holding a smartphone. Key service highlights include consulting a doctor in as quickly as 3 minutes, 4-hour express medicine delivery (marked as new), and GP & Chinese Medicine Practitioner online consultations. A "Experience Now" button with a hand cursor is located at the bottom right.

WDr Watsons Online Doctor

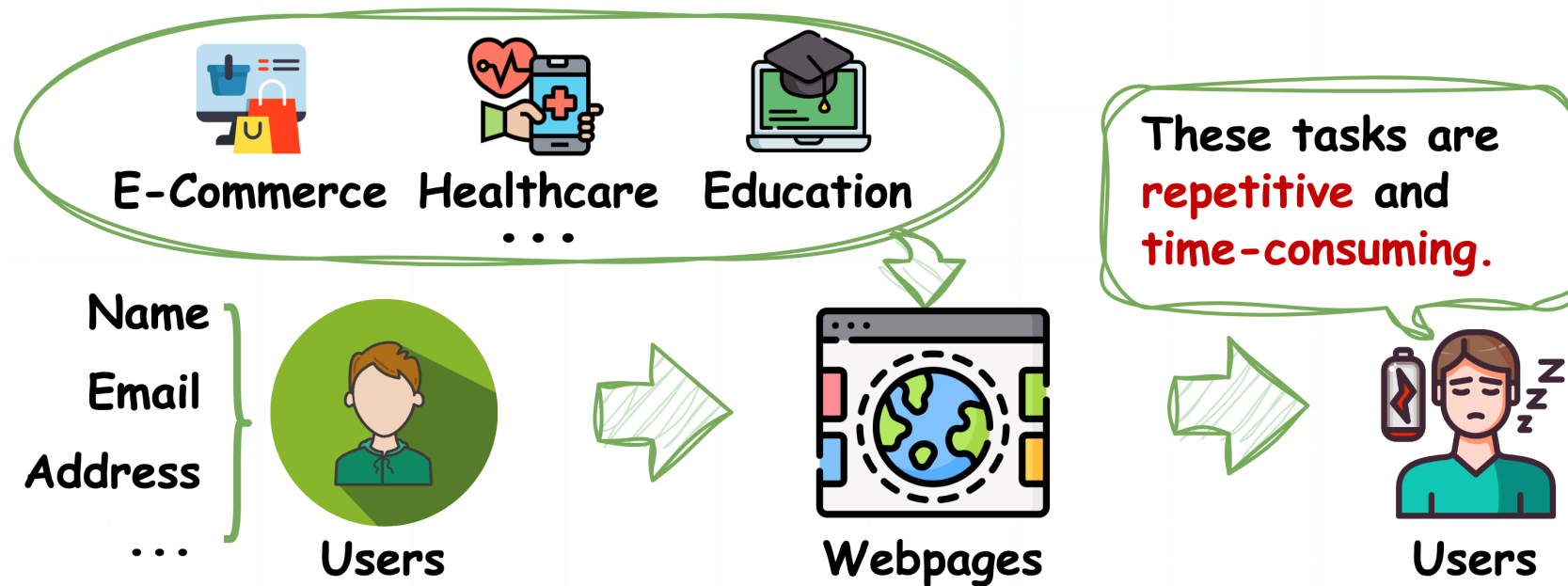
Consult eDoctor Anytime Anywhere

- Consult Doctor in as quickly as 3mins\*
- 4 Hours# Express Medicine Delivery\*\*
- GP & Chinese Medicine Practitioner Online Consultation\*

Experience Now

# Laborious tasks on Web

Despite the importance of the web, many tasks performed on it are **repetitive and time-consuming**, negatively impacting overall quality of life.



# AI Agents for Web

***"GPTs and Assistants are precursors to agents. They will gradually be able to plan and to perform more complex actions on your behalf. These are our first step toward AI Agents"***  
— Sam Altman

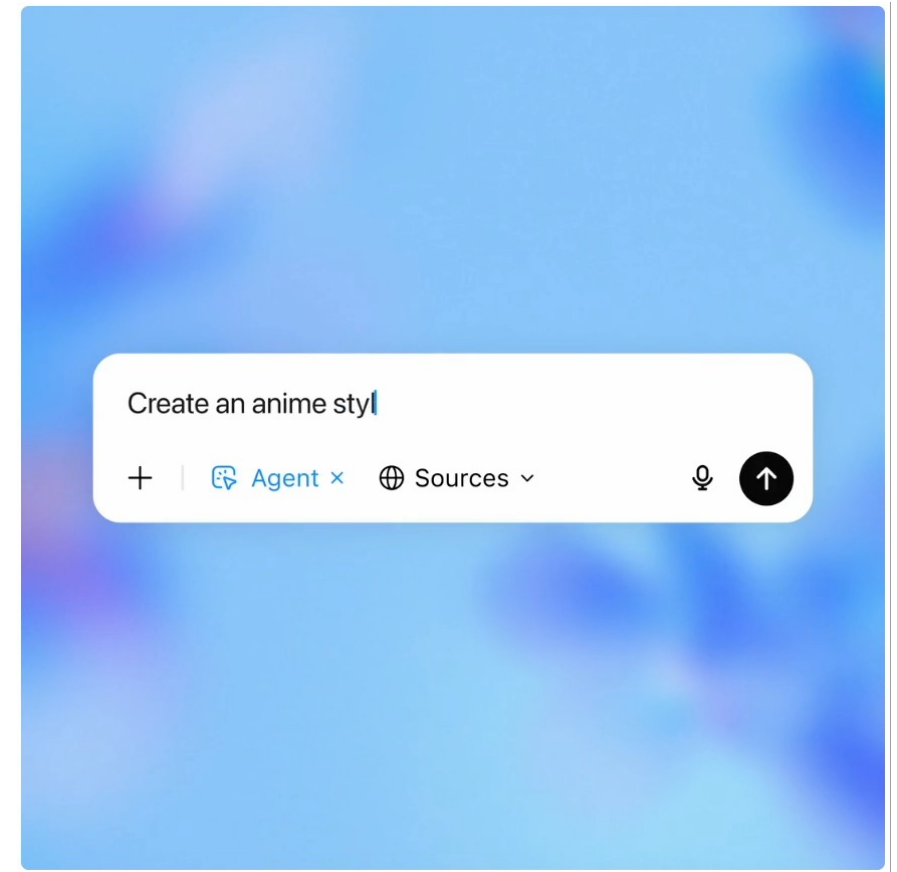
TED Ideas change everything

WATCH D



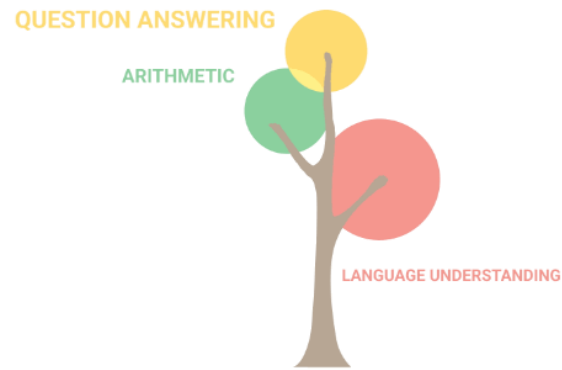
OpenAI's Sam Altman talks ChatGPT, AI agents and superintelligence — live at TED2025

1,997,406 plays | Sam Altman | TED2025 • April 2025



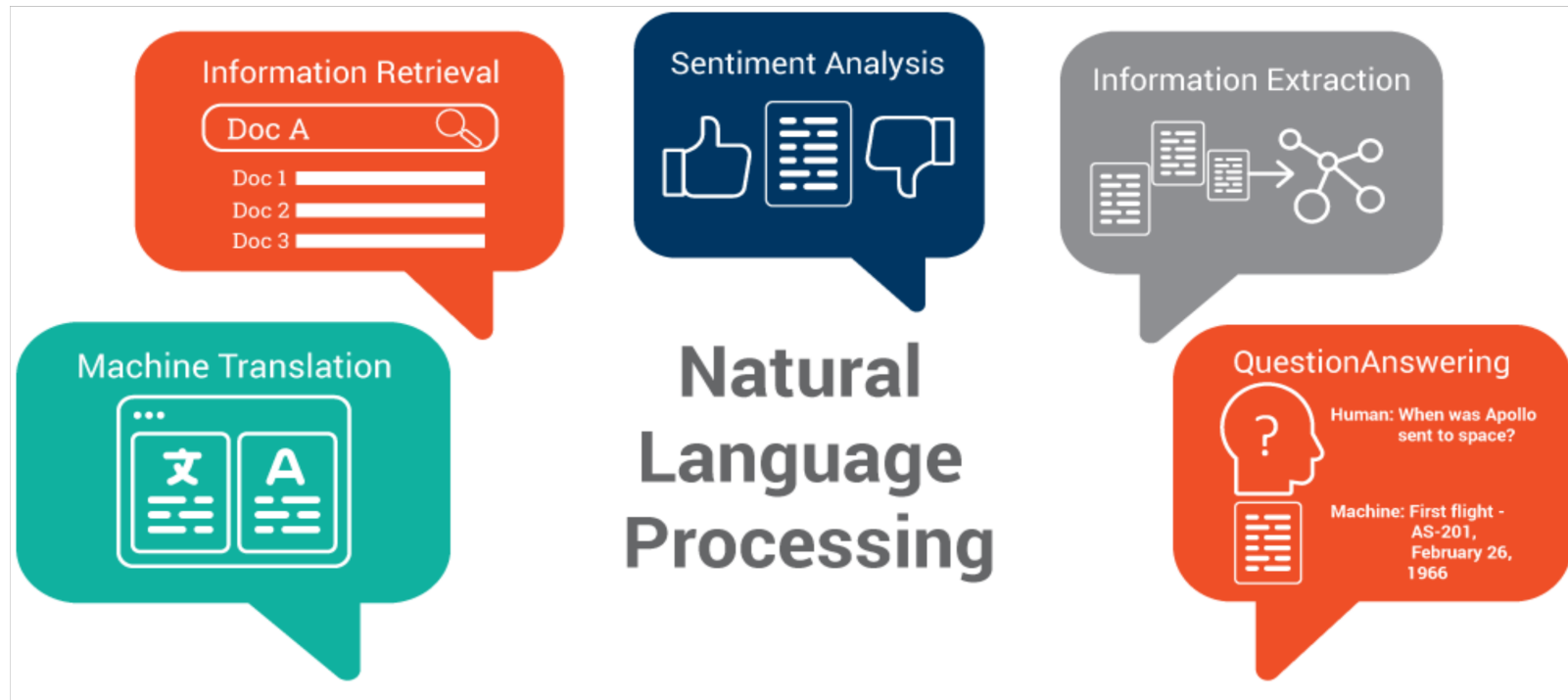
# Large Foundations Models (LFMs)

They Are Changing Our Lives !



8 billion parameters

# LFMs in Natural Language Processing



**Input Text**

A rounded rectangular box containing three horizontal lines representing input text.

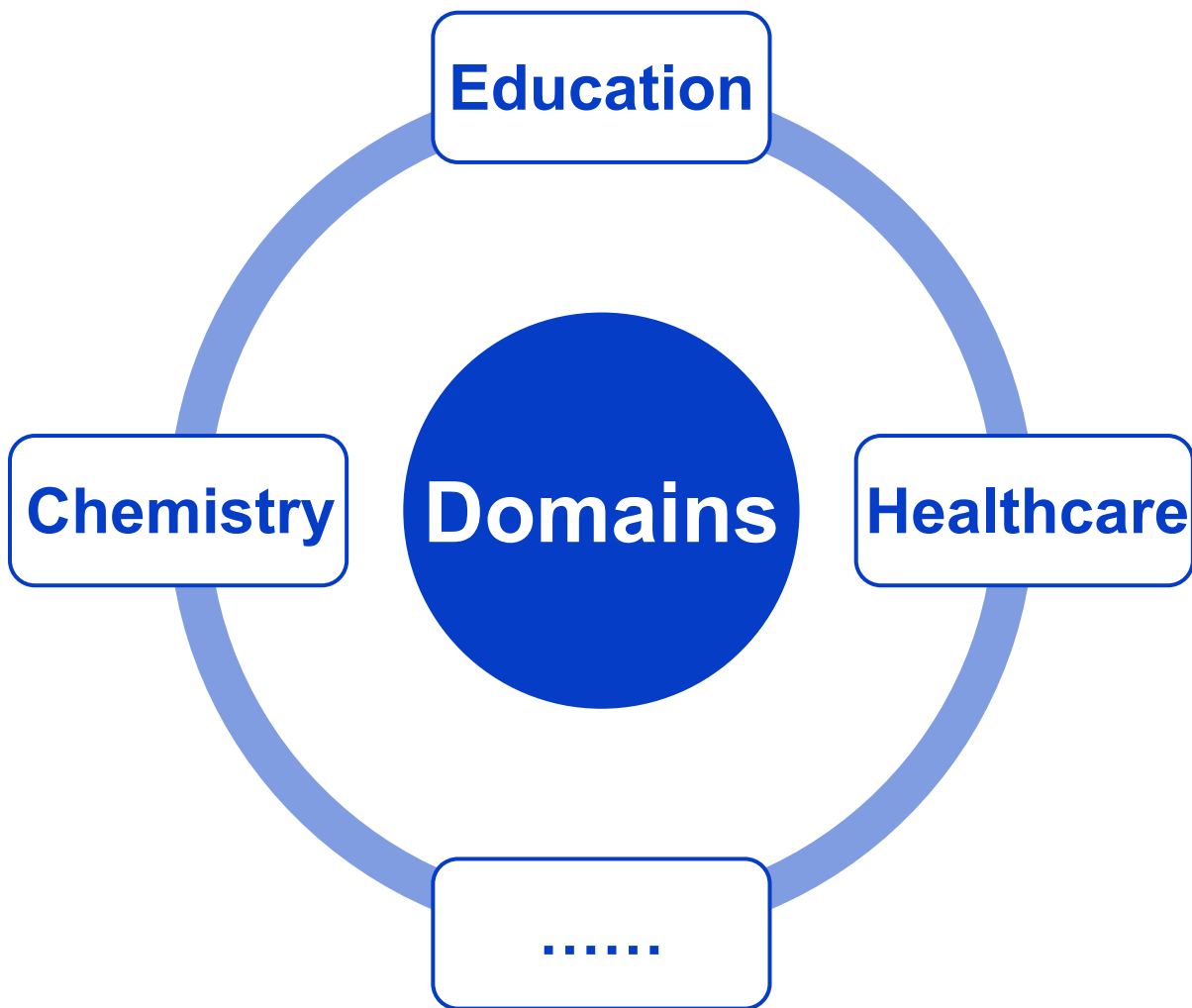


**Generated Text**

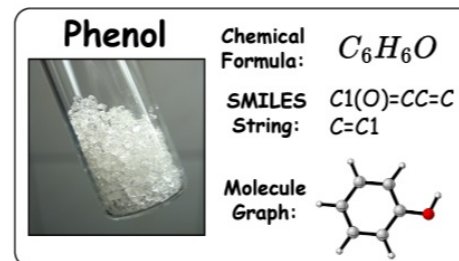
A rounded rectangular box containing three horizontal lines representing generated text.

**Large Language Models (LLMs)**

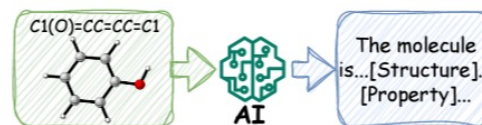
# LFMs in Downstream Domains



## ❑ Molecule discovery, etc.




### (a) Molecule Representations.





### (b) Molecule Captioning.




 ChatGPT

**(a) Molecule Captioning**


Please show me a description of this molecule: "C1=CC=C(C=C1)OC2=CC=CC=C2" 

 The molecule is an aromatic ether in which the oxygen is attached to two phenyl substituents. It has been found in muscat grapes and vanilla. It has a role as a plant metabolite.

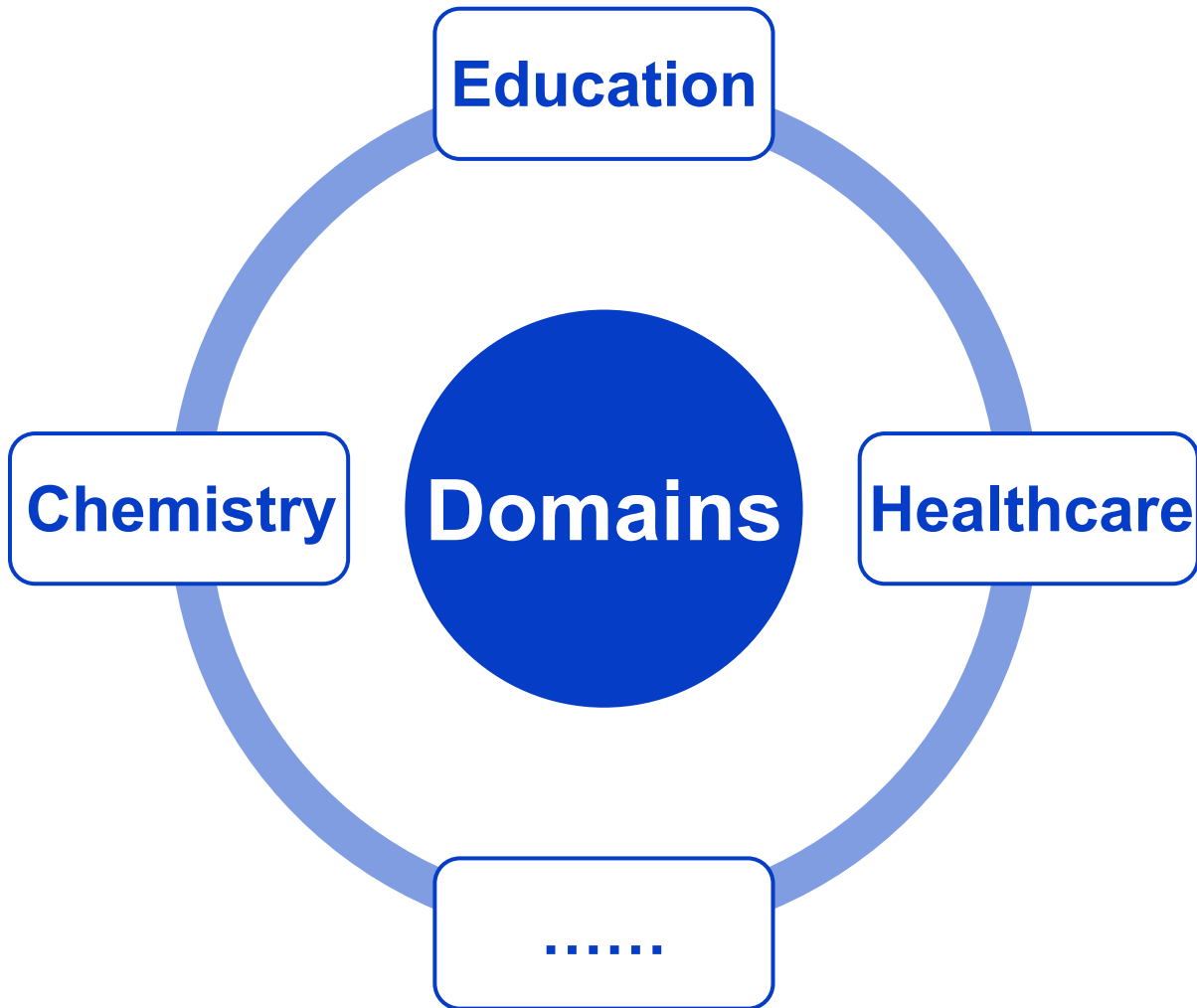
**(b) Text-based Molecule Generation**

Help me generate a molecule based on the given description: 

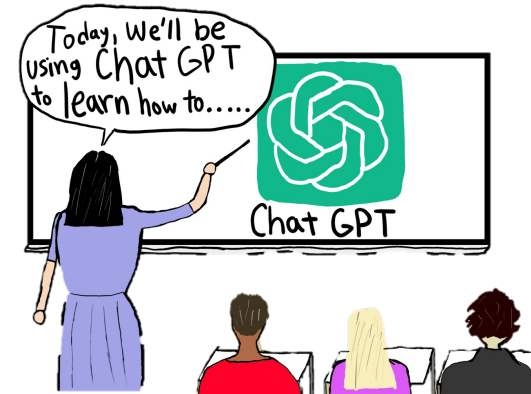
The molecule is a quinolinemonocarboxylate that is the conjugate base of xanthurenic acid, obtained by deprotonation of the carboxy group. It has a role as an animal metabolite. It is a conjugate base of a xanthurenic acid.

 C1=CC2=C(C=C1)[O-]NC(=CC2=O)C(=O)O

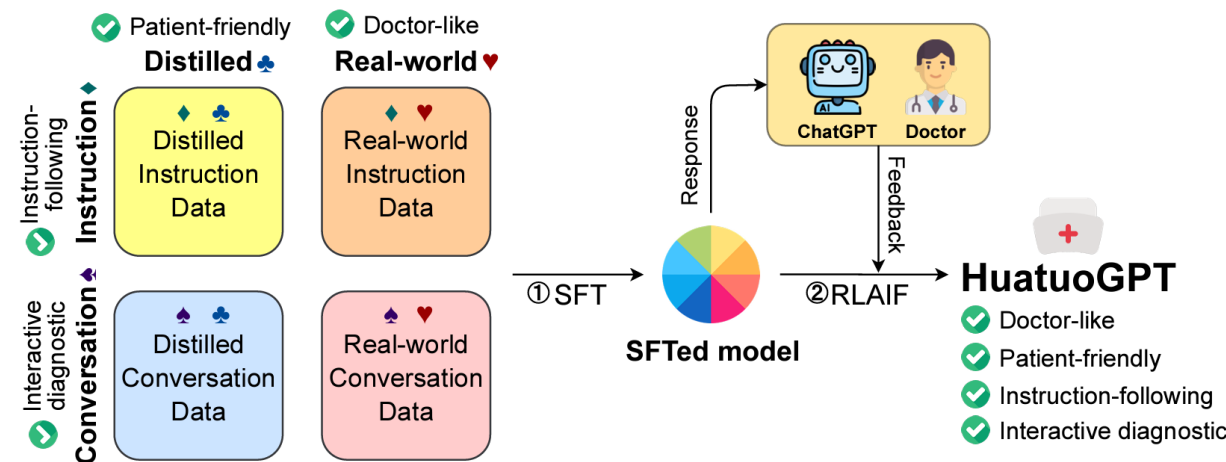
# LFMs in Downstream Domains



## ❑ Curriculum & Teaching, etc.



## ❑ Medical consultation, etc.

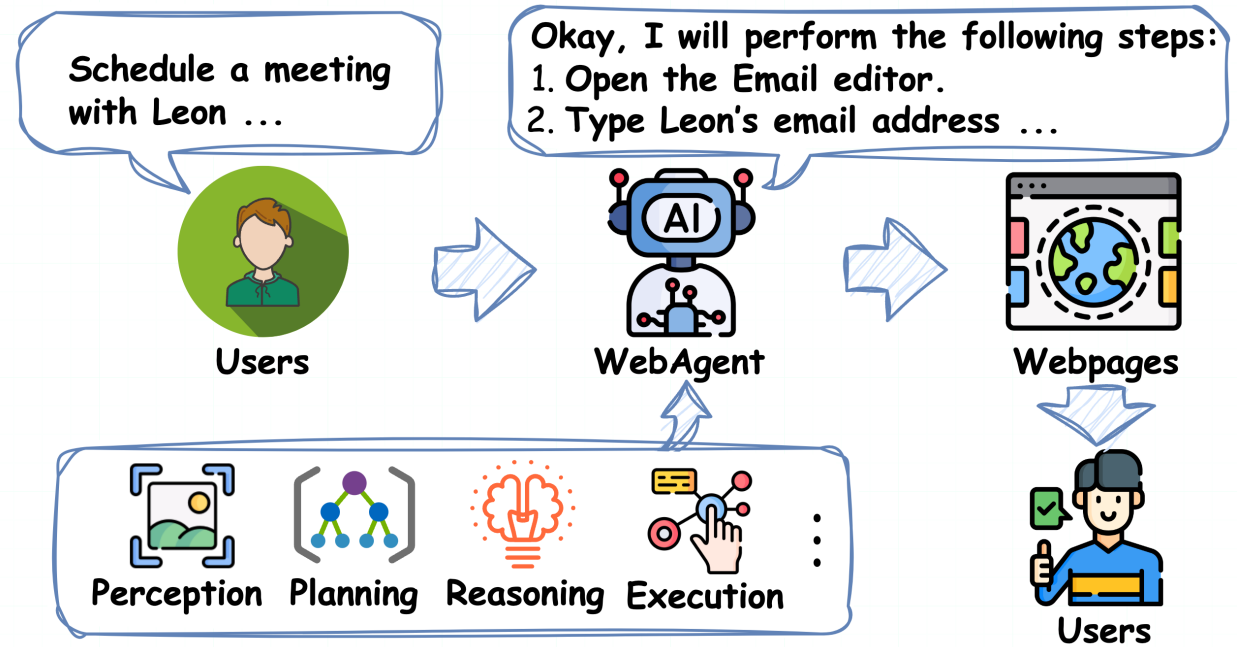


# LFM-Empowered WebAgents

## LFMs:

- *Understanding and Reasoning*
- *Action Generation and Execution*
- *Handing Data with Various Modalities*
- *Memory and Context*

*WebAgents autonomously complete tasks by perceiving the environment, reasoning action sequences, and executing interactions.*



# A Comprehensive Survey Paper

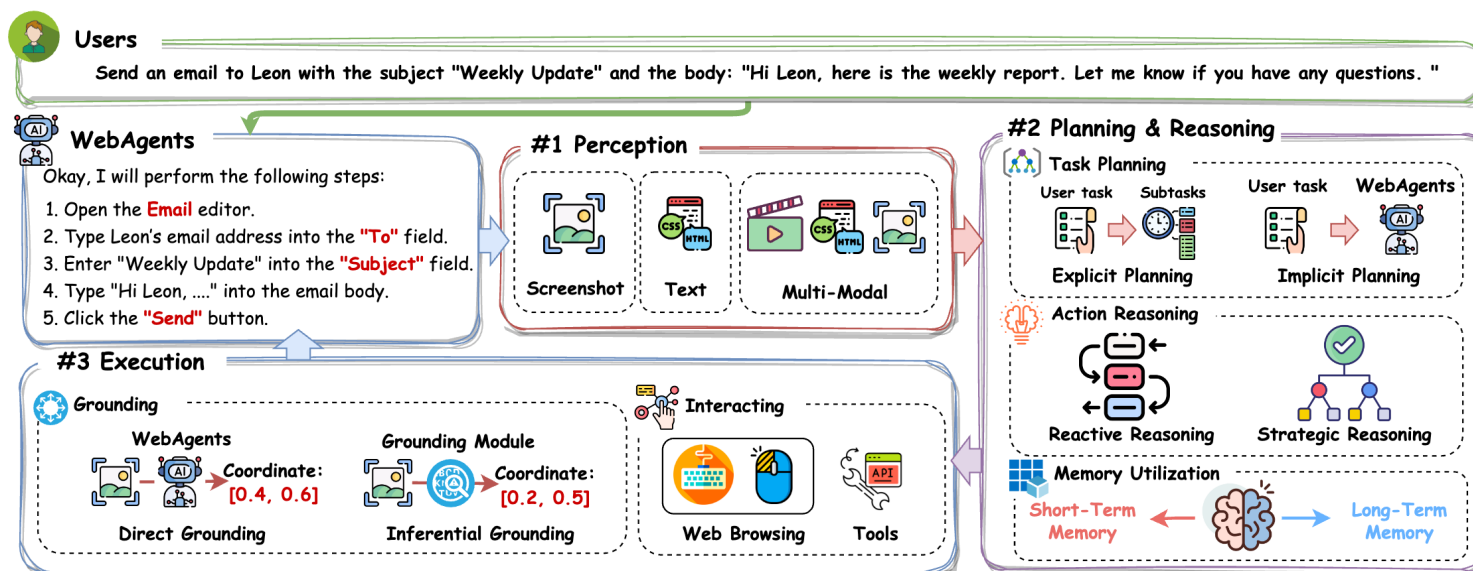
## A Survey of WebAgents: Towards Next-Generation AI Agents for Web Automation with Large Foundation Models

Liangbo Ning<sup>1</sup>, Ziran Liang<sup>1</sup>, Zhuohang Jiang<sup>1</sup>, Haohao Qu<sup>1</sup>, Yujuan Ding<sup>1</sup>,  
Wenqi Fan<sup>1\*</sup>, Xiao-yong Wei<sup>1</sup>, Shanru Lin<sup>2</sup>, Hui Liu<sup>3</sup>, Philip S. Yu<sup>4</sup>, Qing Li<sup>1\*</sup>

<sup>1</sup>The Hong Kong Polytechnic University, <sup>2</sup>City University of Hong Kong,

<sup>3</sup>Michigan State University, <sup>4</sup>University of Illinois at Chicago

<https://arxiv.org/pdf/2503.23350>



Survey paper  
on KDD



Tutorial  
Website (Slides)



Tutorial website: <https://biglemon-ning.github.io/WebAgents/>

# Recruitment

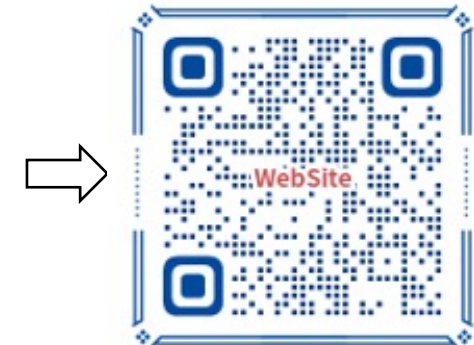
- Our research group are actively recruiting self-motivated **Postdoc, Ph.D. students, and Research Assistants**, etc. **Visiting scholars, interns, and self-funded students** are also welcome. Send me an email if you are interested.
- ❖ Research areas: machine learning (ML), data mining (DM), artificial intelligence (AI), deep learning (DNNs), graph neural networks (GNNs), computer vision (CV), natural language processing (NLP), etc.
- ❖ Position Details:  
<https://wenqifano3.github.io/openings.html>



# Tutorial Outline

- ⦿ Part 1: Introduction of WebAgents
- ⦿ **Part 2: Preliminaries of AI Agents and LFM-based WebAgents**
- Part 3: Architectures of WebAgents
- Coffee Break
- Part 4: Training of WebAgents
- Part 5: Trustworthy WebAgents
- Part 6: Future directions of WebAgents

Website of this tutorial  
Check out the slides and more information!



# PART 2: Preliminaries of AI Agents and LFM-based WebAgents

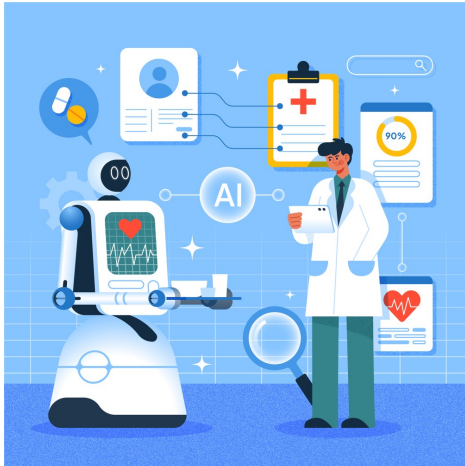


Website of this tutorial

- RL-based Agents
  - Overview of RL-based Agents
  - Deep Reinforcement Learning (DRL)
  - Applications and Challenges
- LFM-empowered Agents
  - Overview of LFM-empowered Agents
  - Applications and Challenges
- AI Agents for Web Automation
  - Preliminaries of WebAgents
  - Applications and Challenges

# Preliminaries of AI Agents and LFM-based WebAgents

- ❑ AI Agents with human-like reasoning and autonomous decision-making capabilities have revolutionized various domains.
  - Reinforcement learning (RL)-based Agents
  - LFM-empowered Agents.



Medicine



Finance



Education

# PART 2: Preliminaries of AI Agents and LFM-based WebAgents

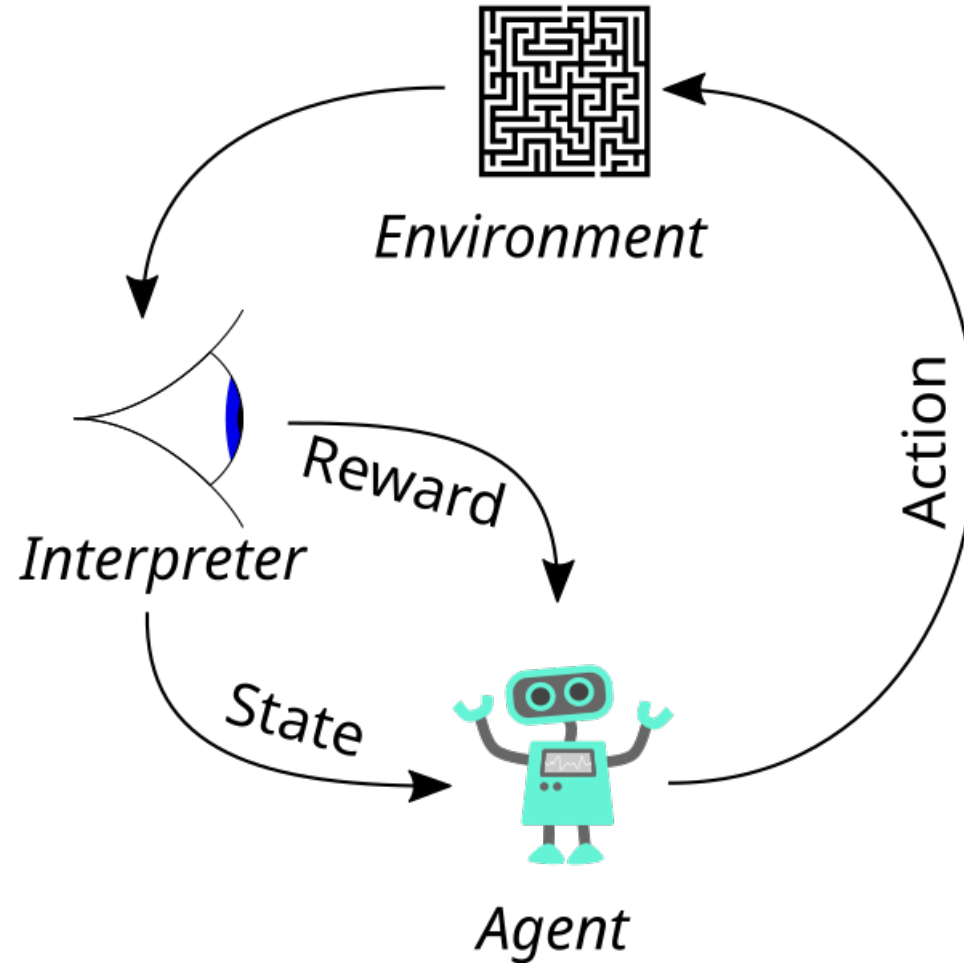


Website of this tutorial

- **RL-based Agents**
  - Overview of RL-based Agents
  - Deep Reinforcement Learning (DRL)
  - Applications and Challenges
- LFM-empowered Agents
  - Overview of LFM-empowered Agents
  - Applications and Challenges
- AI Agents for Web Automation
  - Preliminaries of WebAgents
  - Applications and Challenges

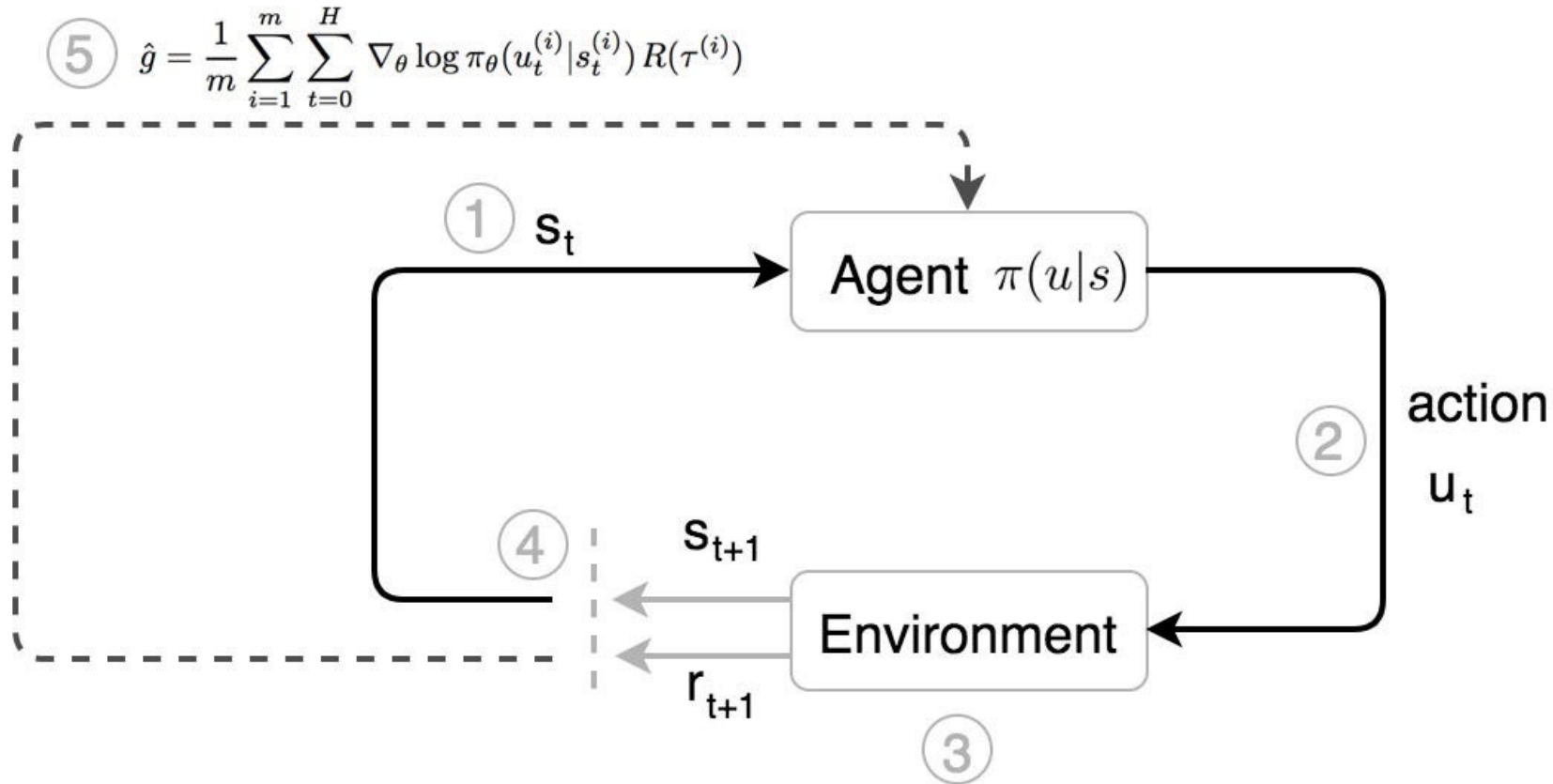
# Reinforcement Learning

- ❑ **Q-Learning** learns the optimal action-value function by iteratively updating Q-values based on rewards and future estimates.



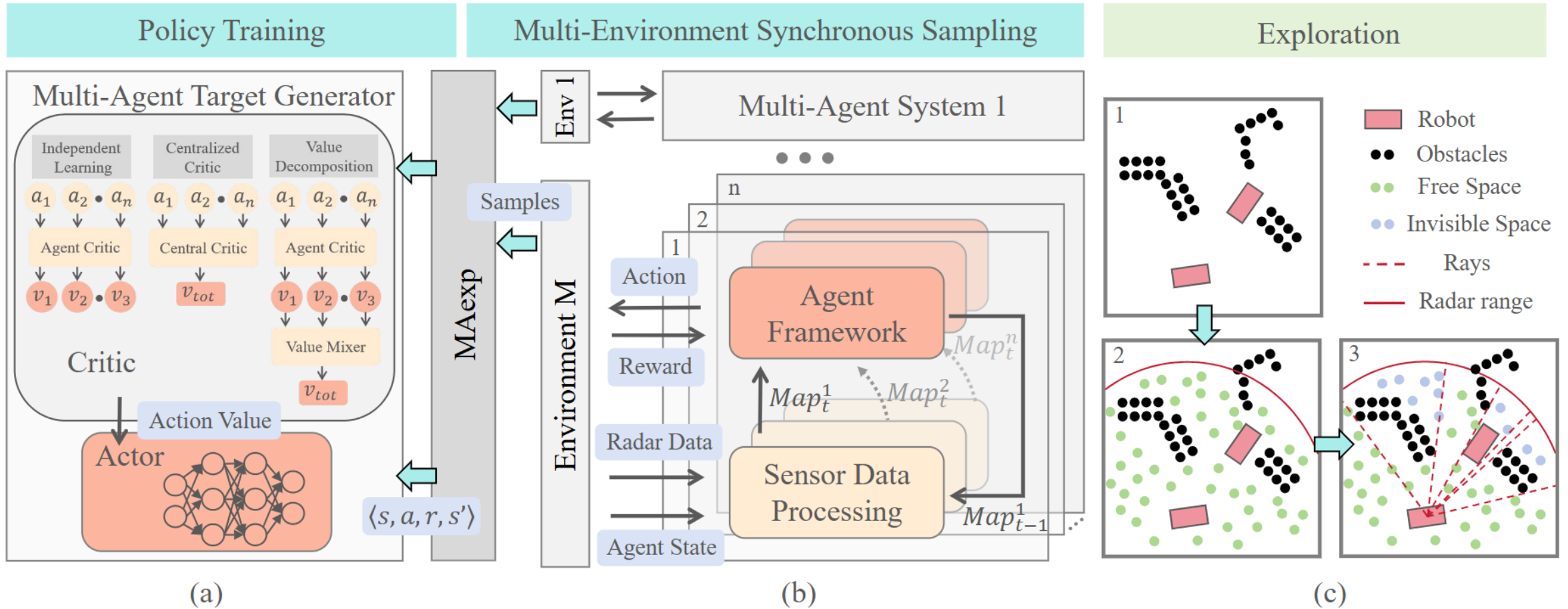
# Reinforcement Learning

- **Policy Gradient** directly optimizes actions by maximizing expected rewards.



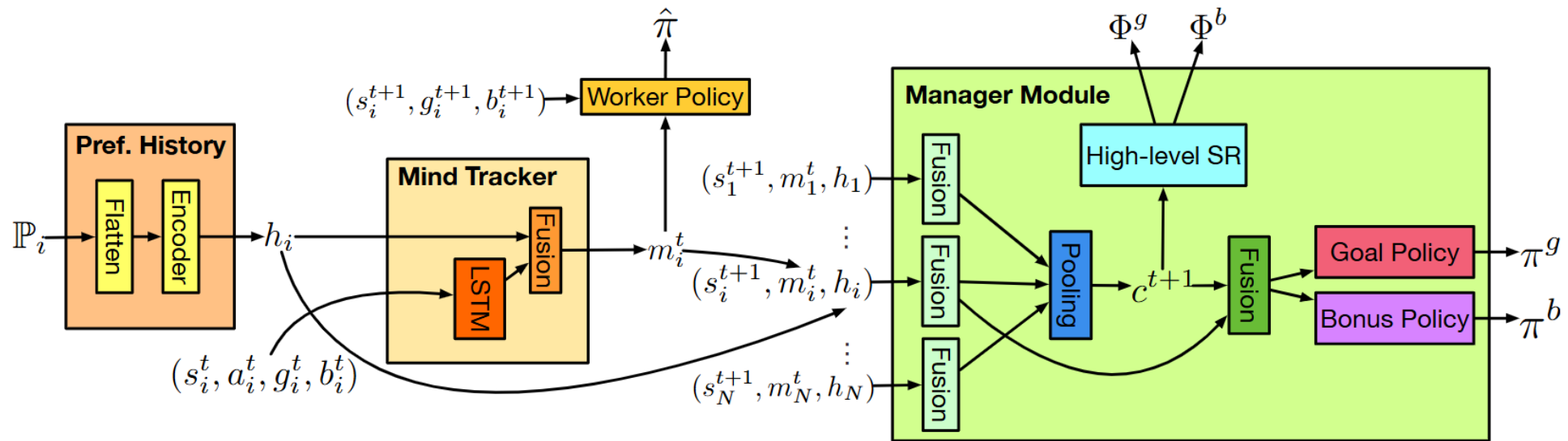
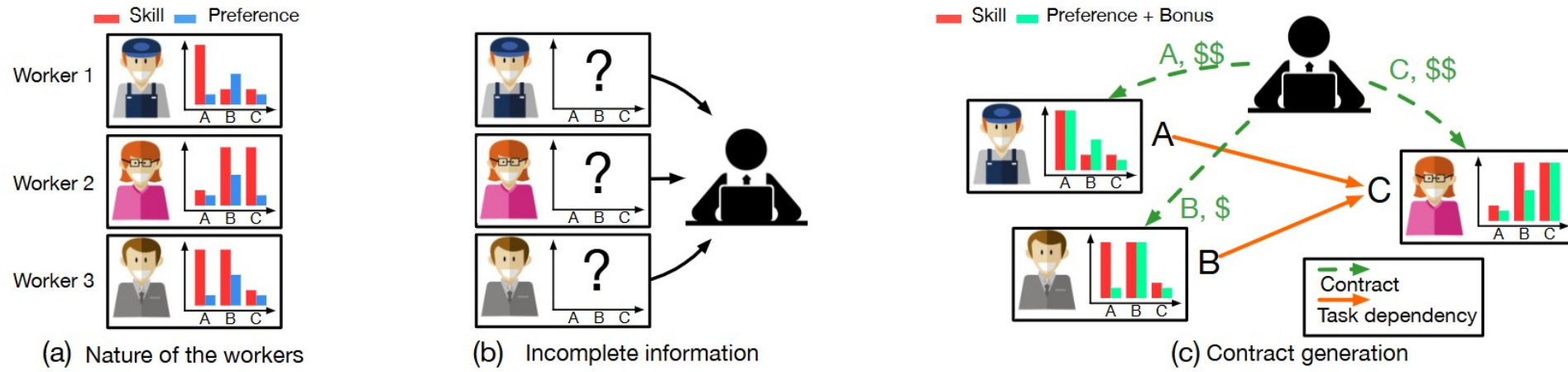
# RL-based Agents

## □ Exploration (Maexp)



# RL-based Agents

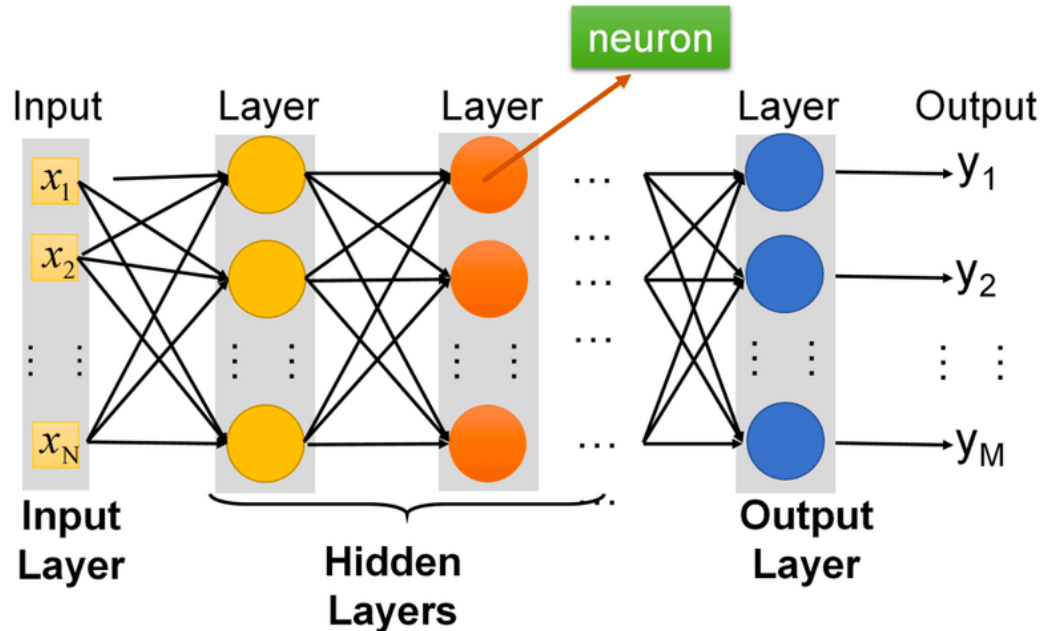
## Management (M<sub>3</sub>RL)



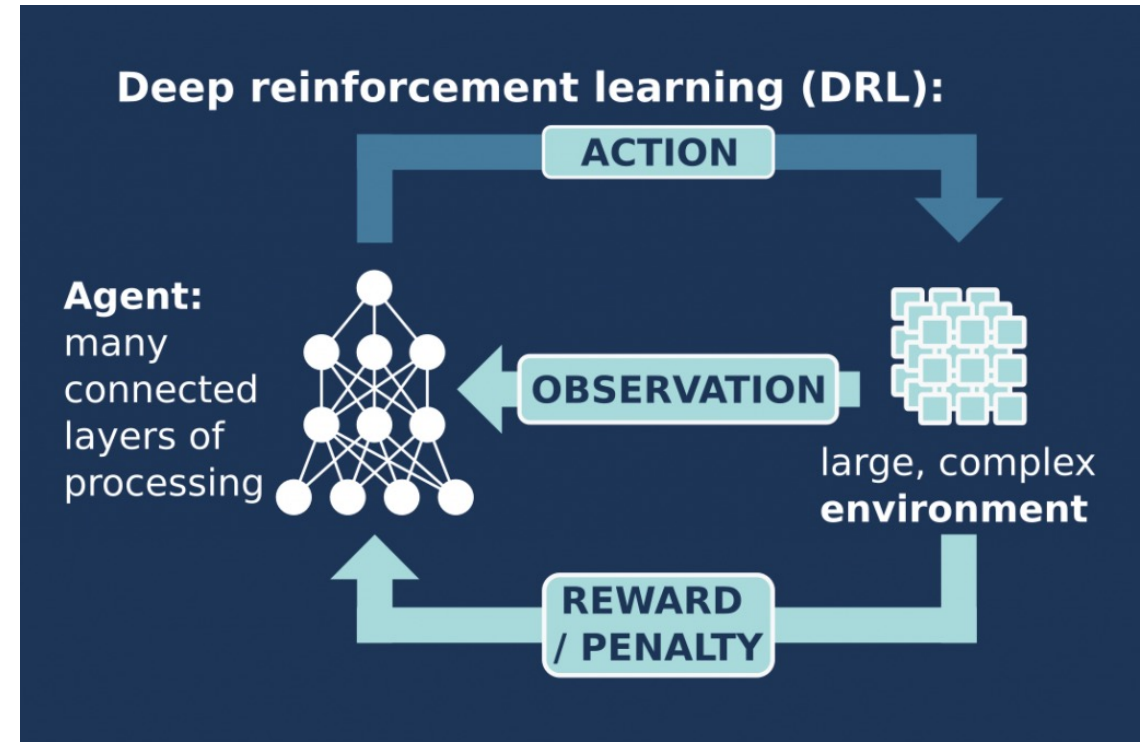
# Deep Reinforcement Learning

- The success of deep learning has led to **Deep Reinforcement Learning** (DRL), combining neural networks with reinforcement learning for complex decision-making.

Deep Neural Network

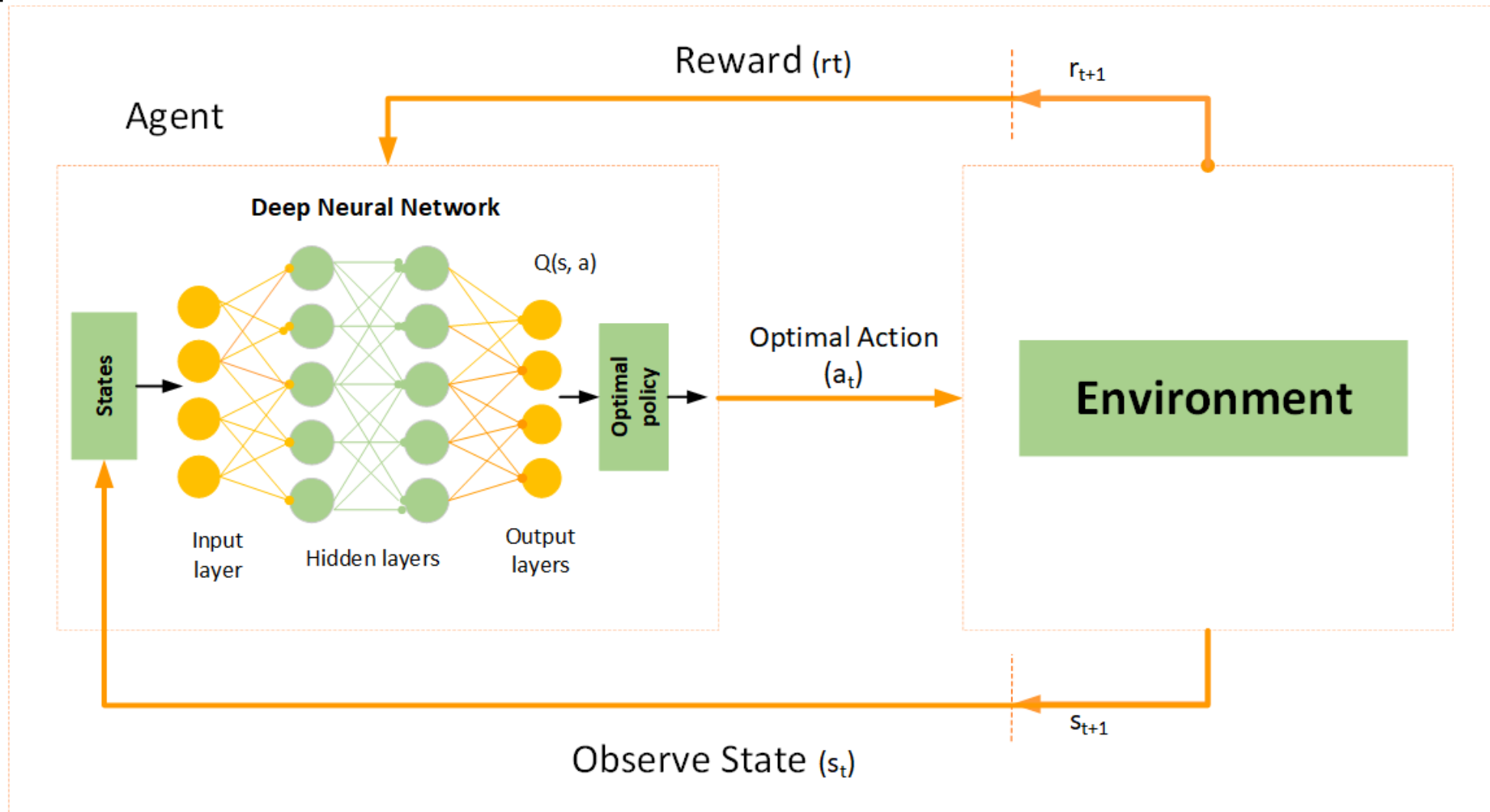


Deep Reinforcement Learning



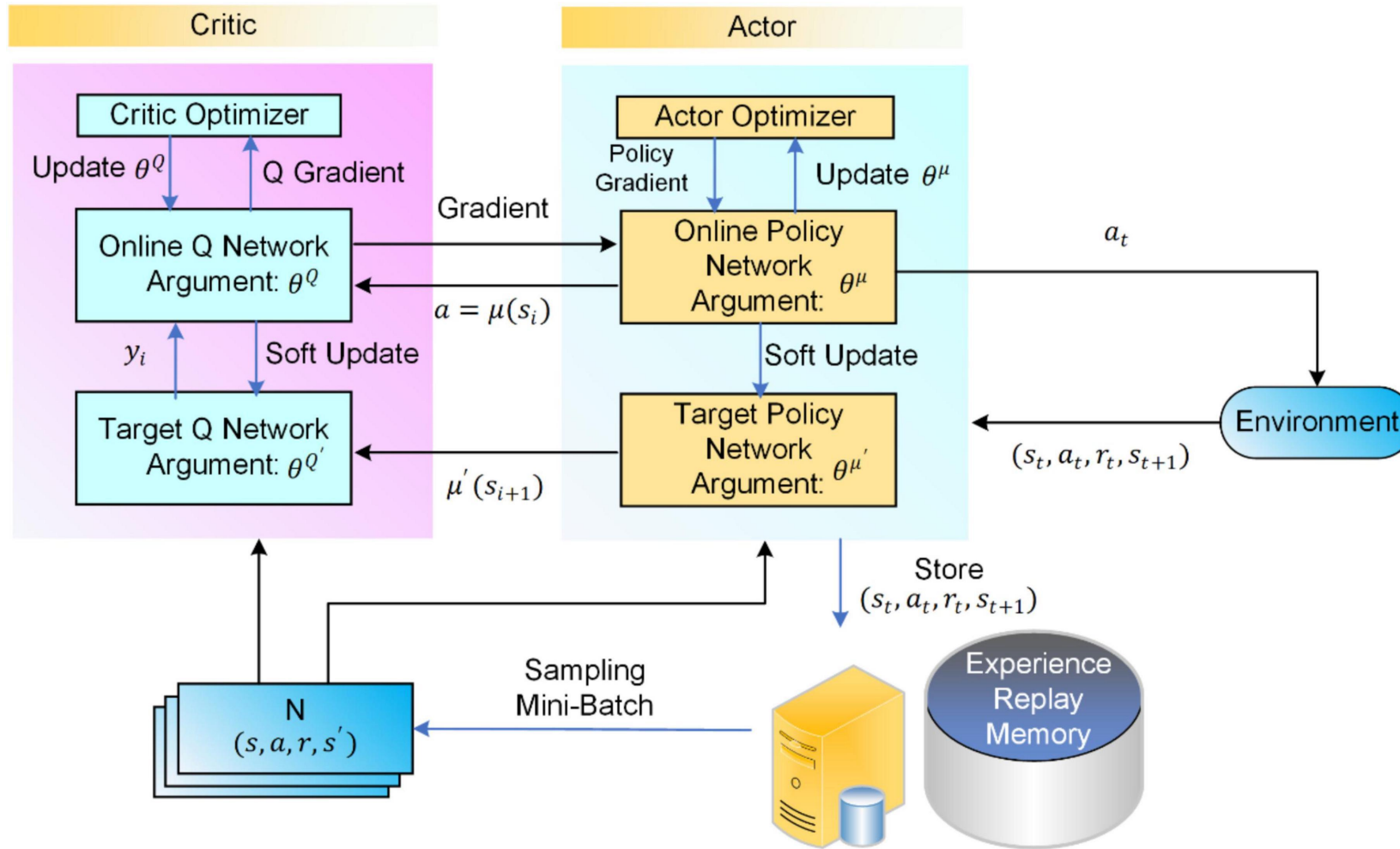
# Deep Reinforcement Learning

## □ Deep Q Network



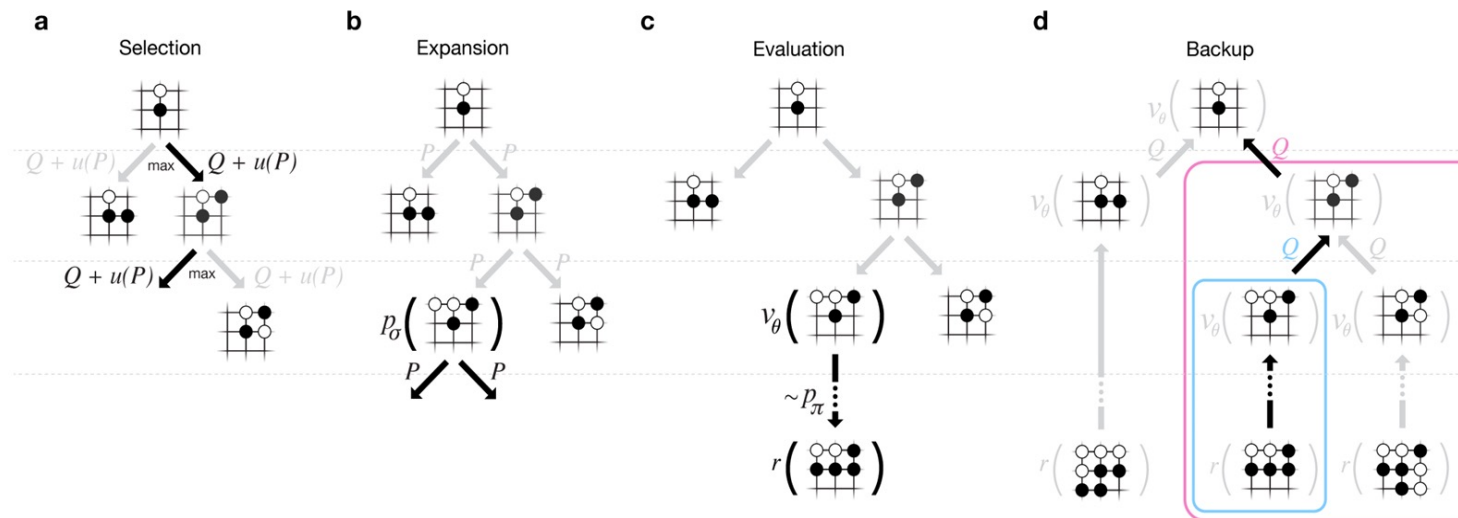
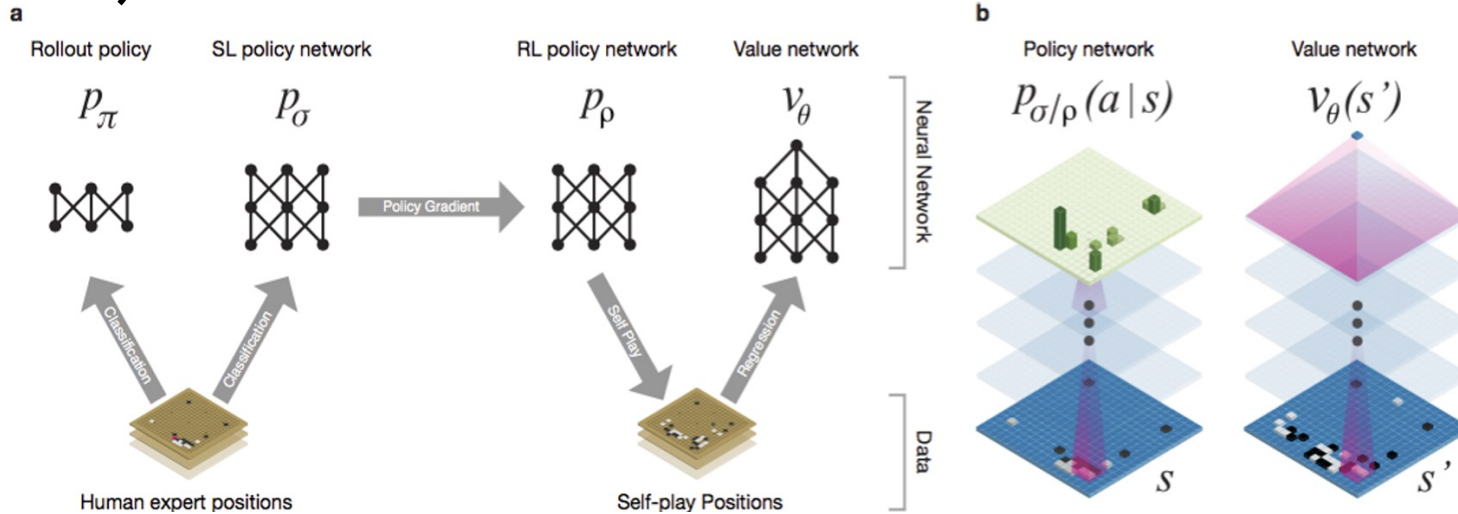
# Deep Reinforcement Learning

## □ Deep Deterministic Policy Gradient



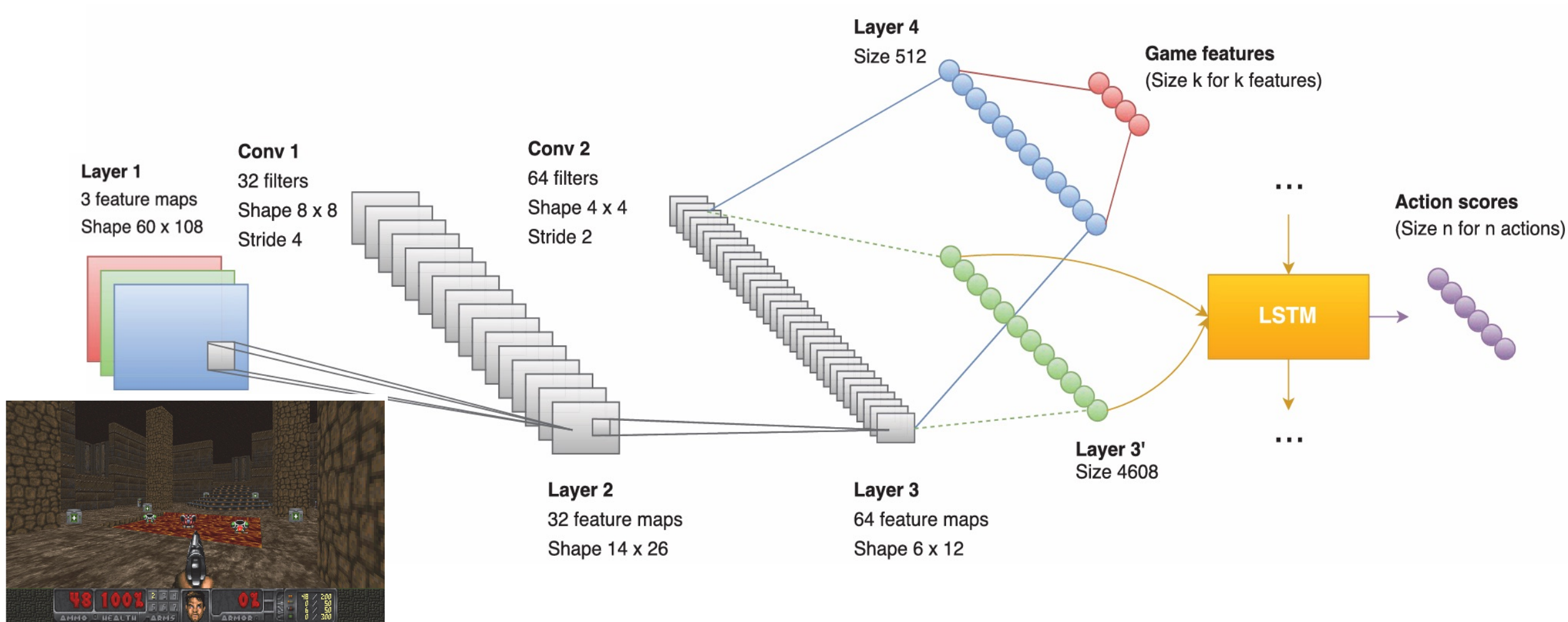
# Deep Reinforcement Learning Agents

## Game (AlphaGo)



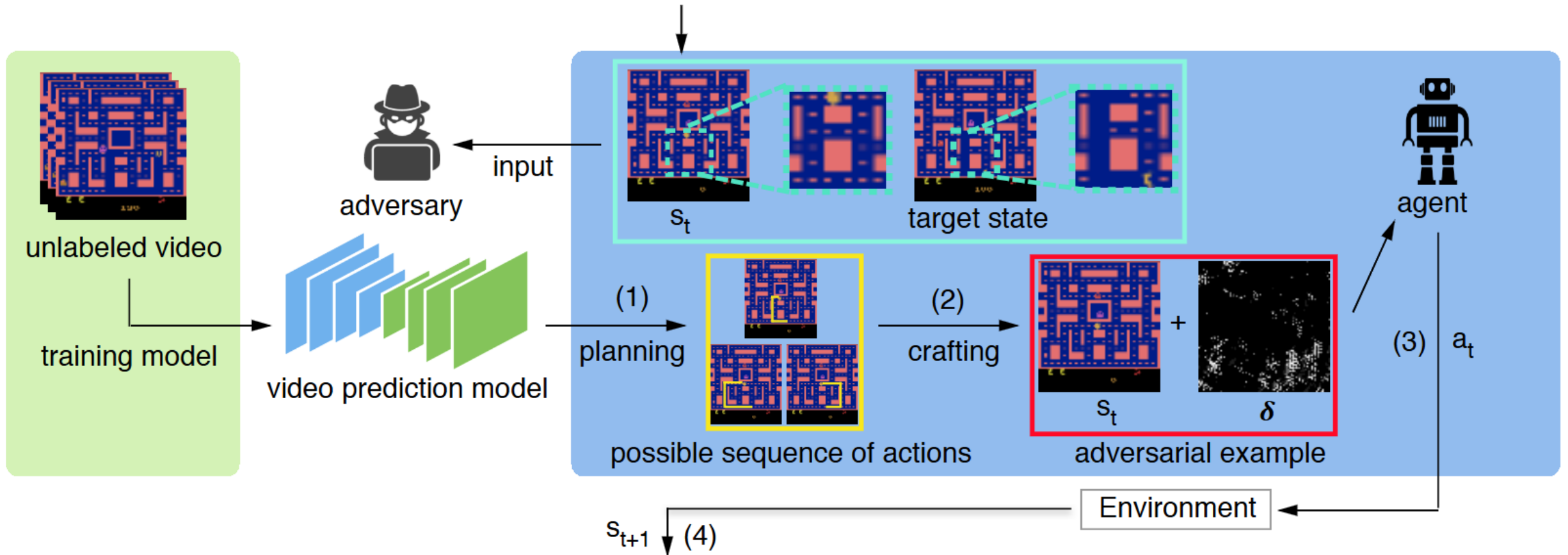
# Deep Reinforcement Learning Agents

## Game (Playing FPS Games with Deep Reinforcement Learning)



# Deep Reinforcement Learning Agents

## Attack (Tactics of Adversarial Attack on Deep Reinforcement Learning Agents)



# Applications and Challenges

## ☐ Applications

- ✔ ➤ Game (Planning)
- ✔ ➤ Finance (Decision-making)
- ✔ ➤ ...

## ☐ Challenges

- ✘ ➤ Lack of World Knowledge
- ✘ ➤ Lack of Adaptability
- ✘ ➤ ...



Go games



Finance

# PART 2: Preliminaries of AI Agents and LFM-based WebAgents

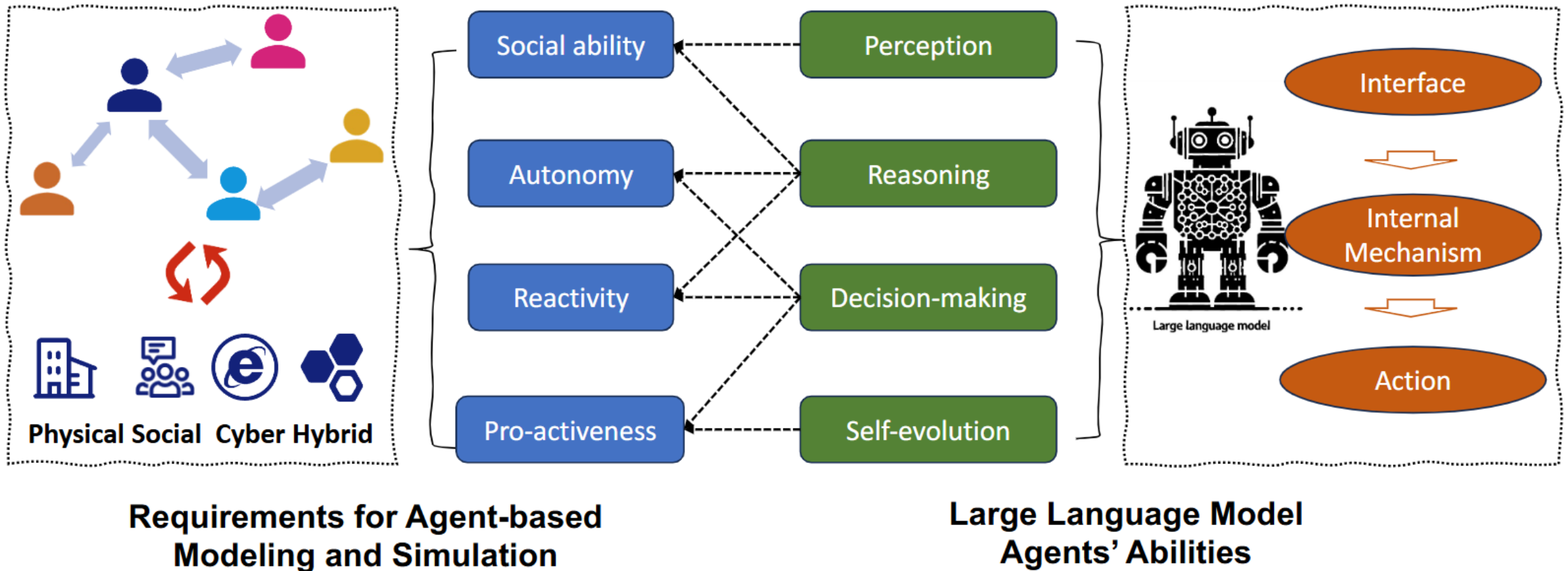


Website of this tutorial

- RL-based Agents
  - Overview of RL-based Agents
  - Deep Reinforcement Learning (DRL)
  - Applications and Challenges
- ⦿ **LFM-empowered Agents**
  - **Overview of LFM-empowered Agents**
  - **Applications and Challenges**
- AI Agents for Web Automation
  - Preliminaries of WebAgents
  - Applications and Challenges

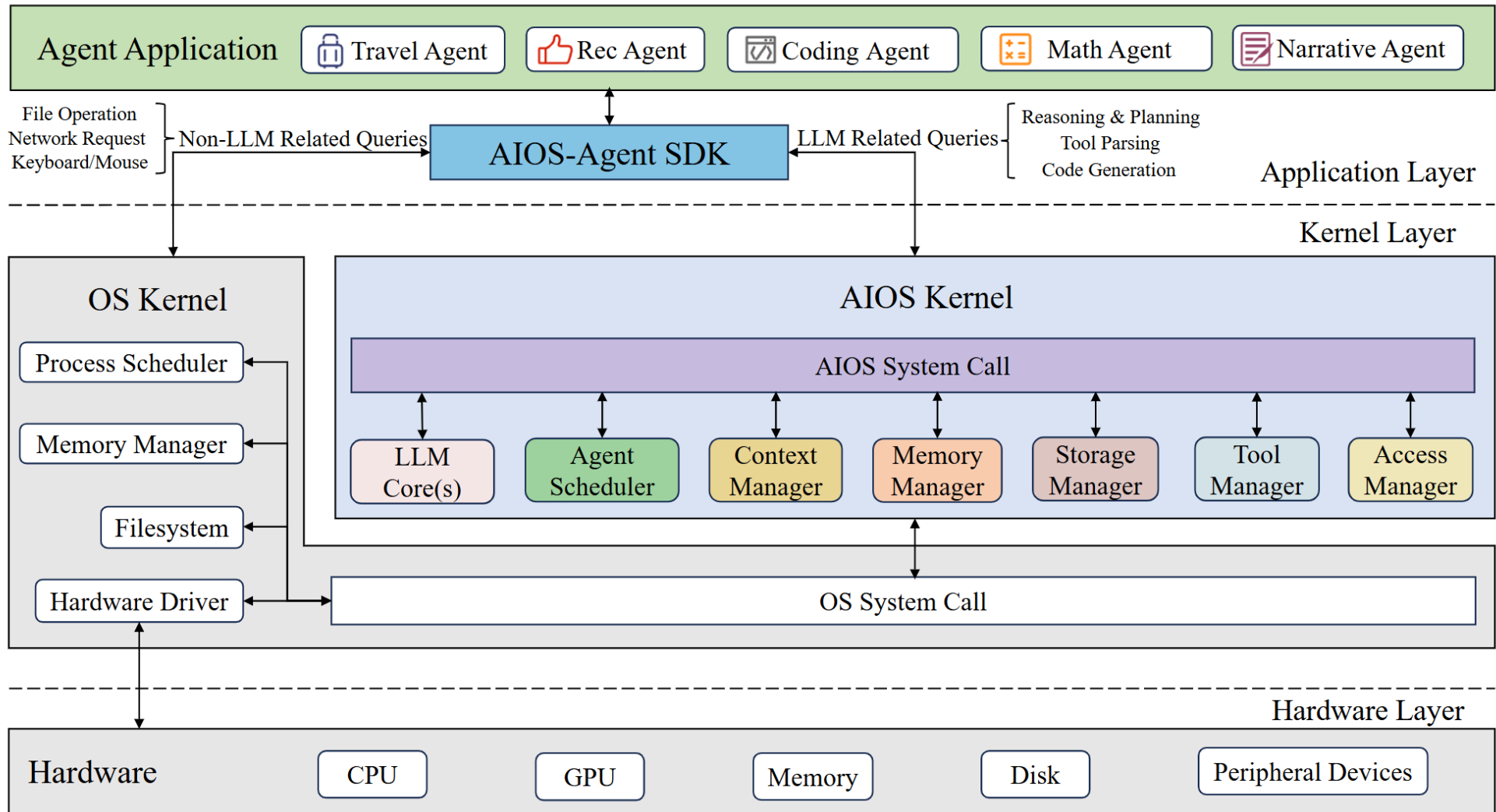
# Overview of LFM-empowered Agents

- ❑ **Large Foundation Models** with billion-level parameters have demonstrated remarkable intelligence characterized by rich intrinsic knowledge.



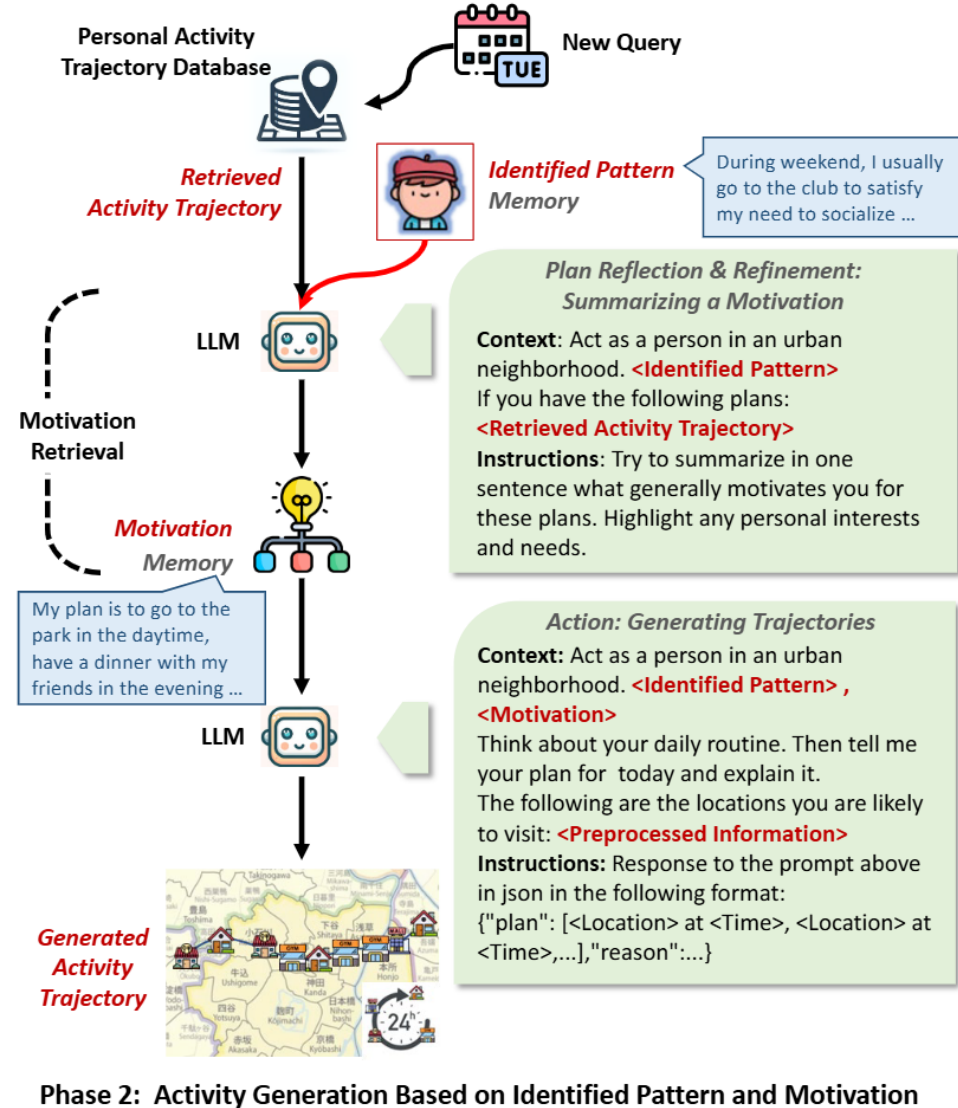
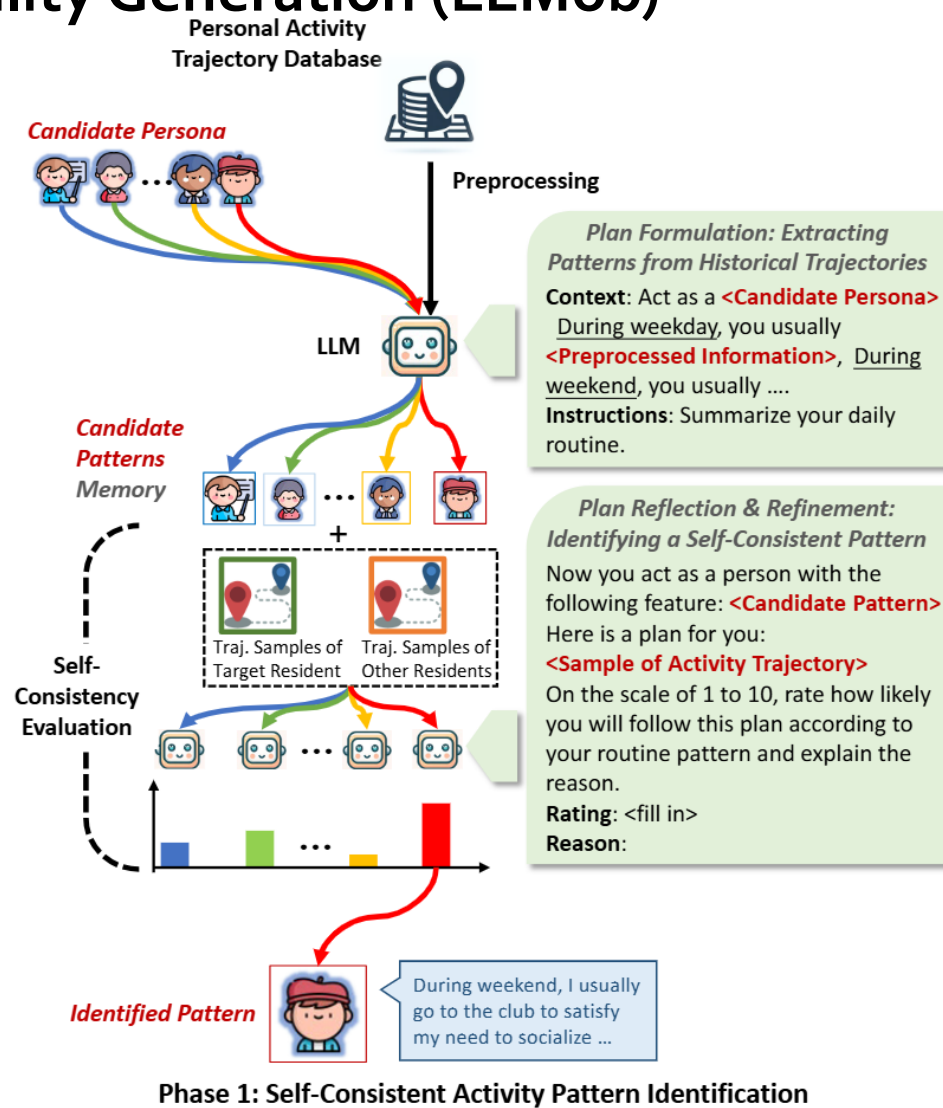
# Applications and Challenges

## □ OS (AIOS)



# Applications and Challenges

## □ Mobility Generation (LLMob)



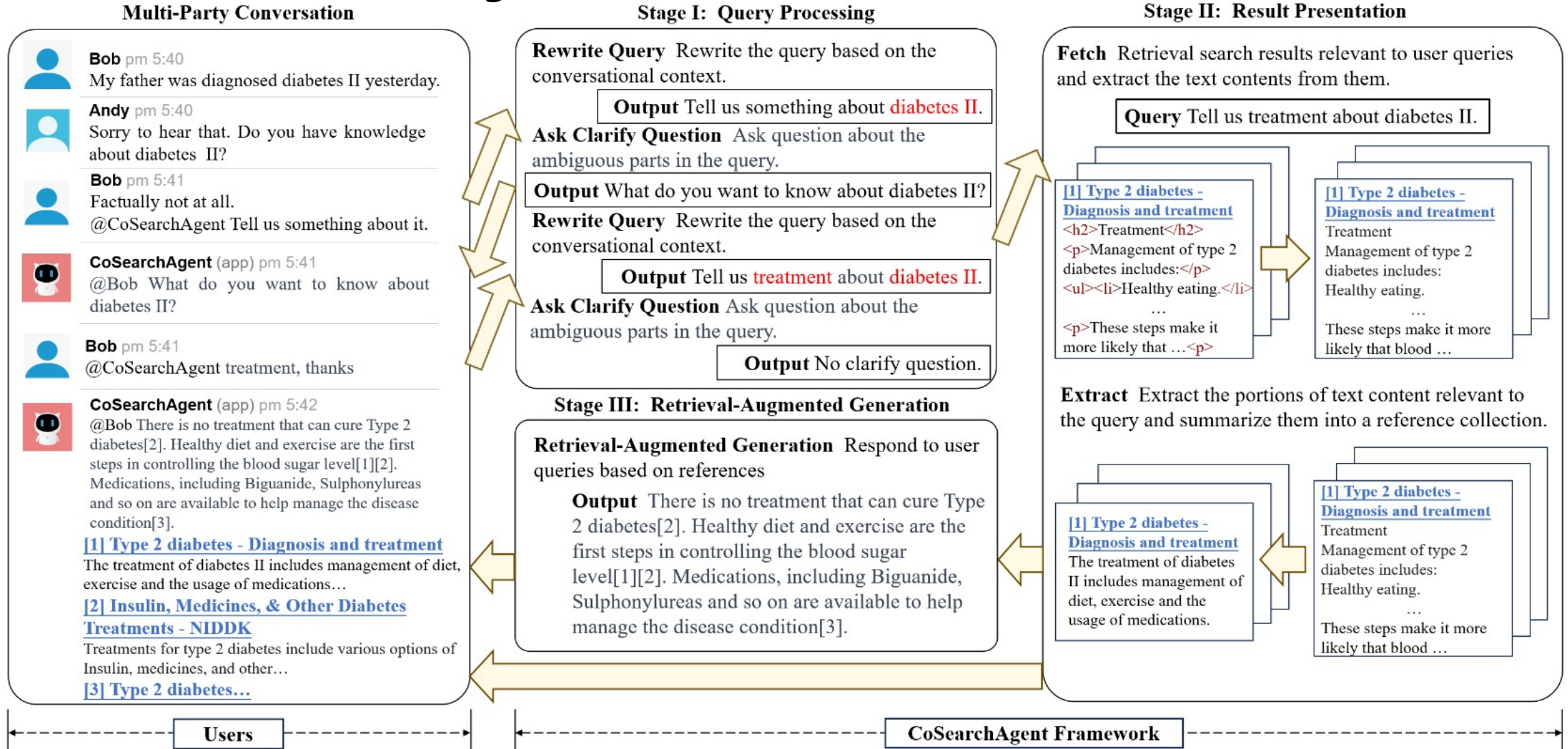
# Applications and Challenges

## ❑ World Model (WorldCoder)



# Applications and Challenges

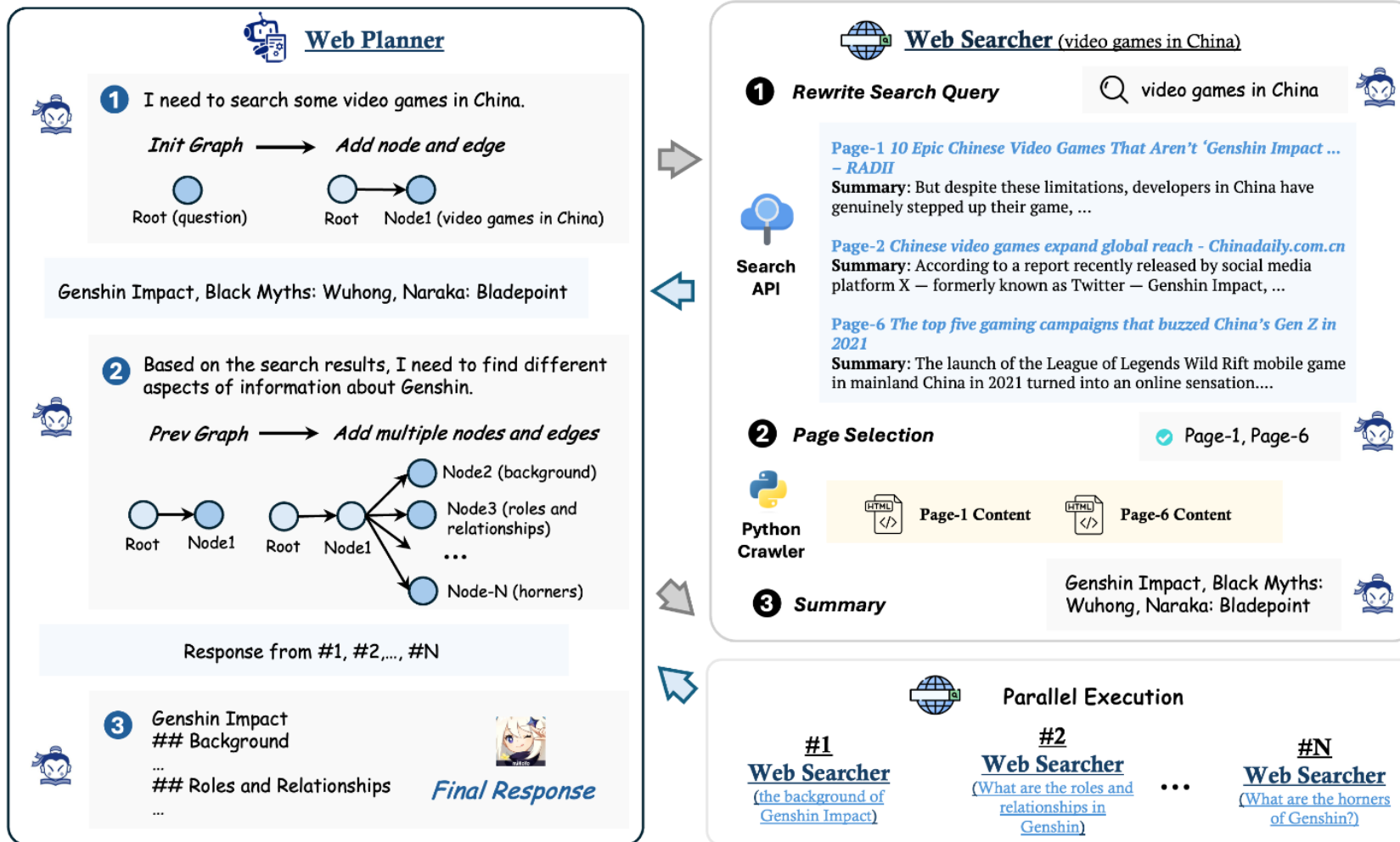
## Web Search (CoSearchAgent)



# Applications and Challenges

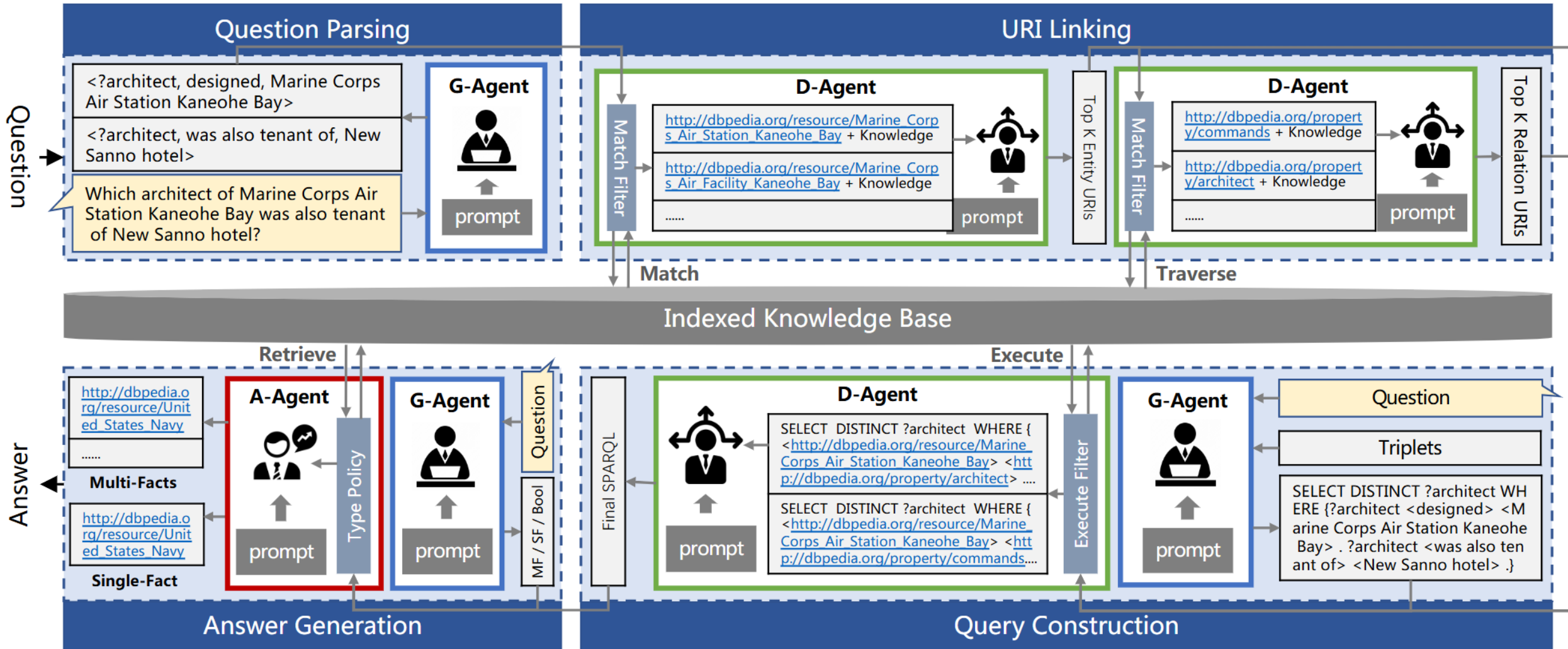
## ❑ Web Search (MindSearch)

**?** Question: Could you introduce one video game produced in China with details?



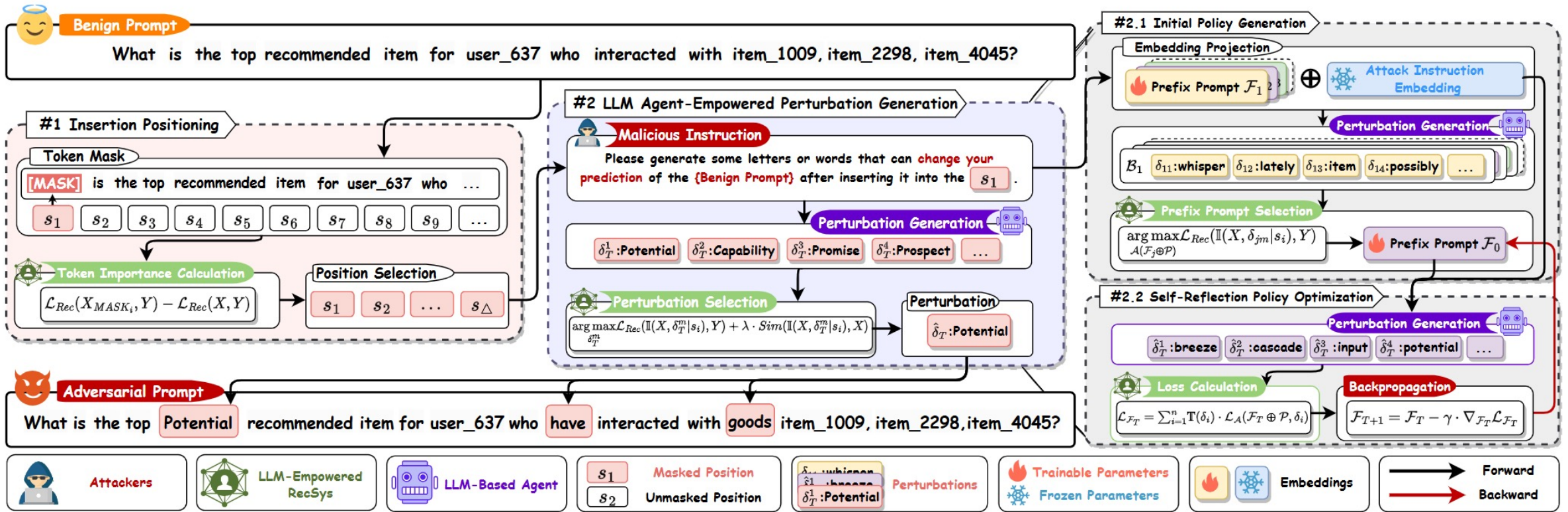
# Applications and Challenges

## KBQA (Triad)



# Applications and Challenges

## RecSys Attack (CheatAgent)



# PART 2: Preliminaries of AI Agents and LFM-based WebAgents

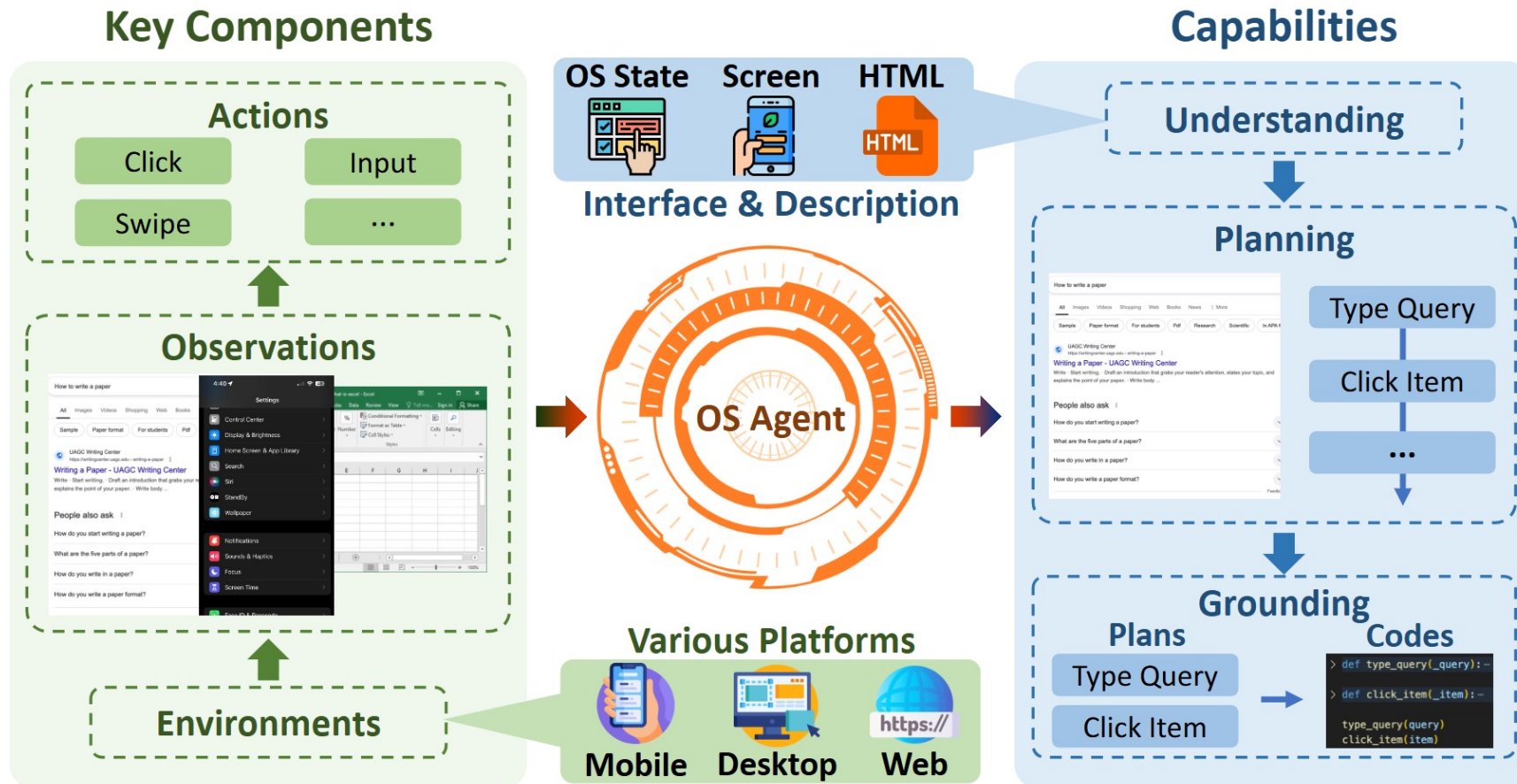


Website of this tutorial

- RL-based Agents
  - Overview of RL-based Agents
  - Deep Reinforcement Learning (DRL)
  - Applications and Challenges
- LFM-empowered Agents
  - Overview of LFM-empowered Agents
  - Applications and Challenges
- ⊙ **AI Agents for Web Automation**
  - **Preliminaries of WebAgents**

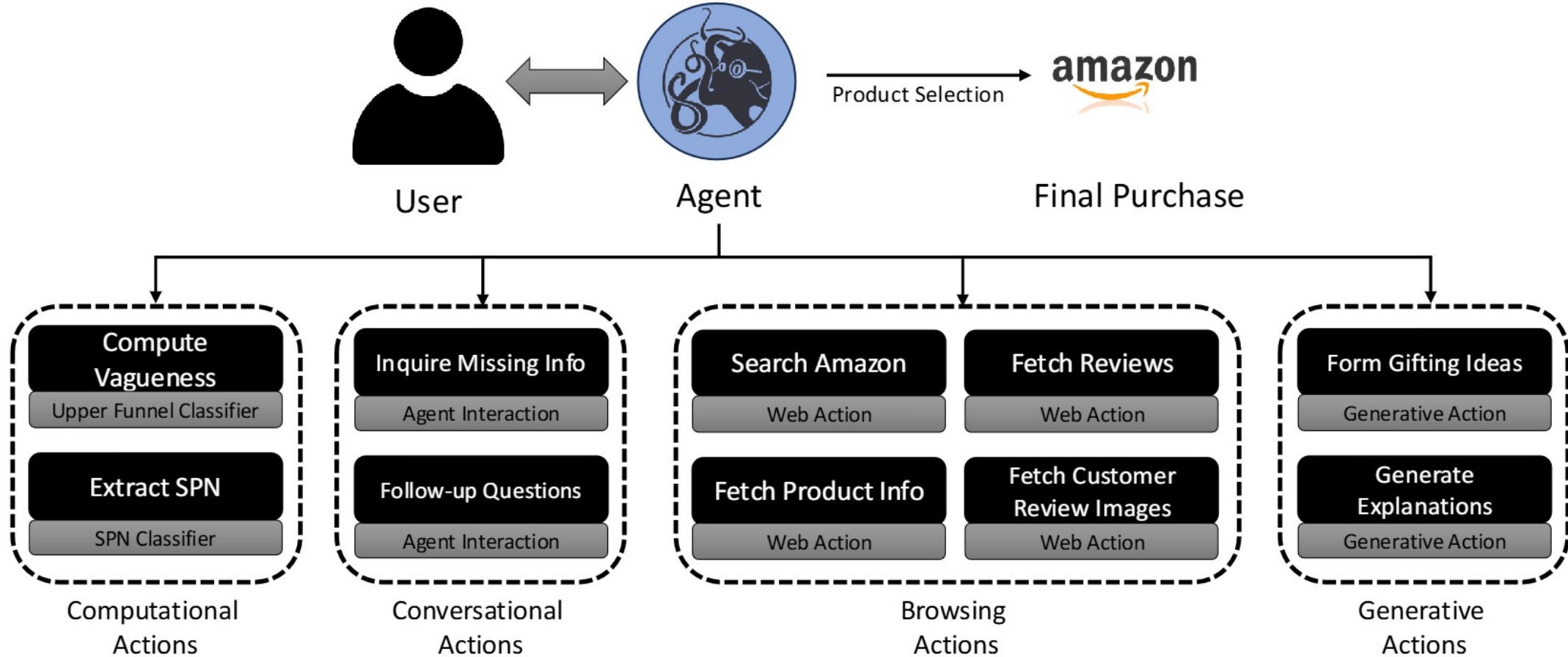
# Preliminaries of WebAgents

- WebAgents aim to perceive the environment and reason over user instructions to generate and execute actions step-by-step toward fulfilling the user's goal.



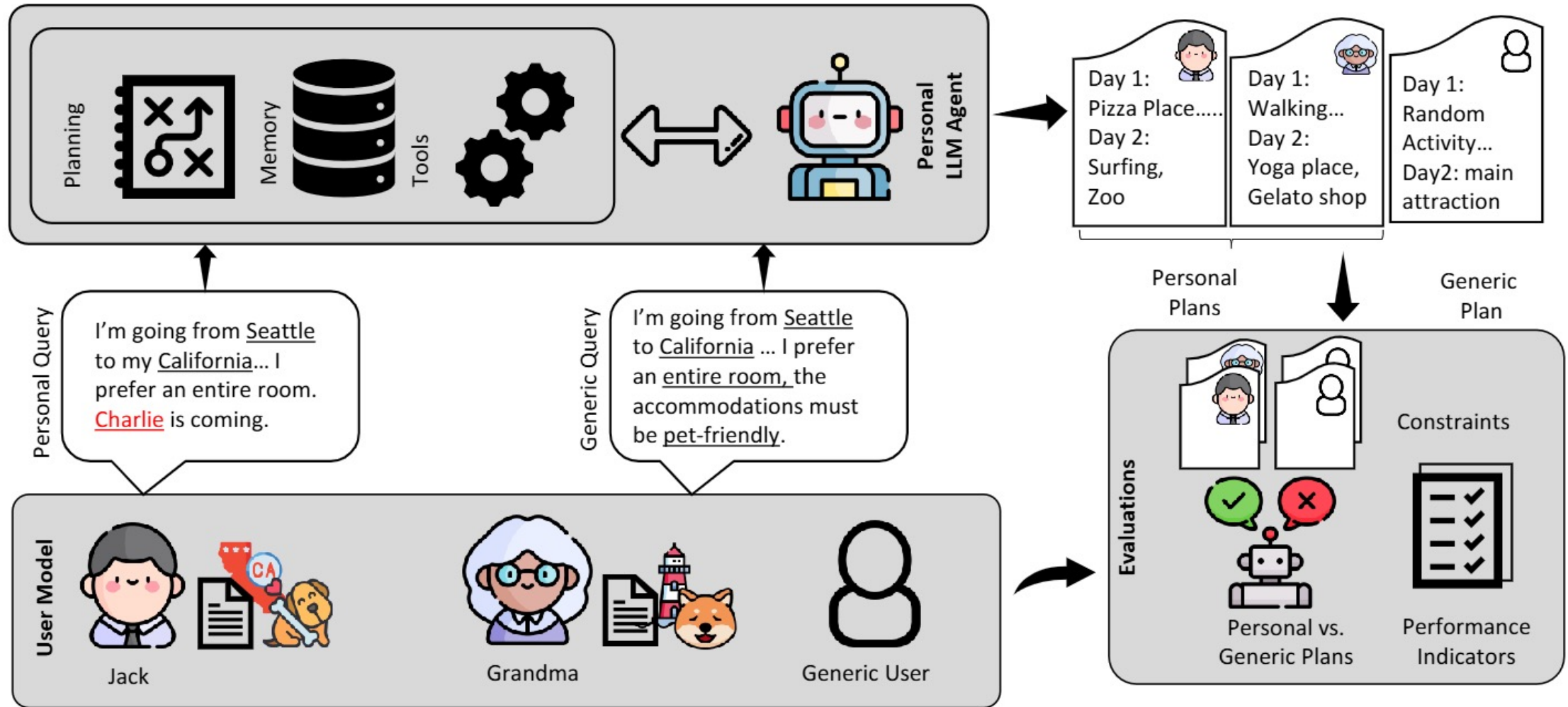
# Applications and Challenges

## Shopping (SPN Shopping Agent)



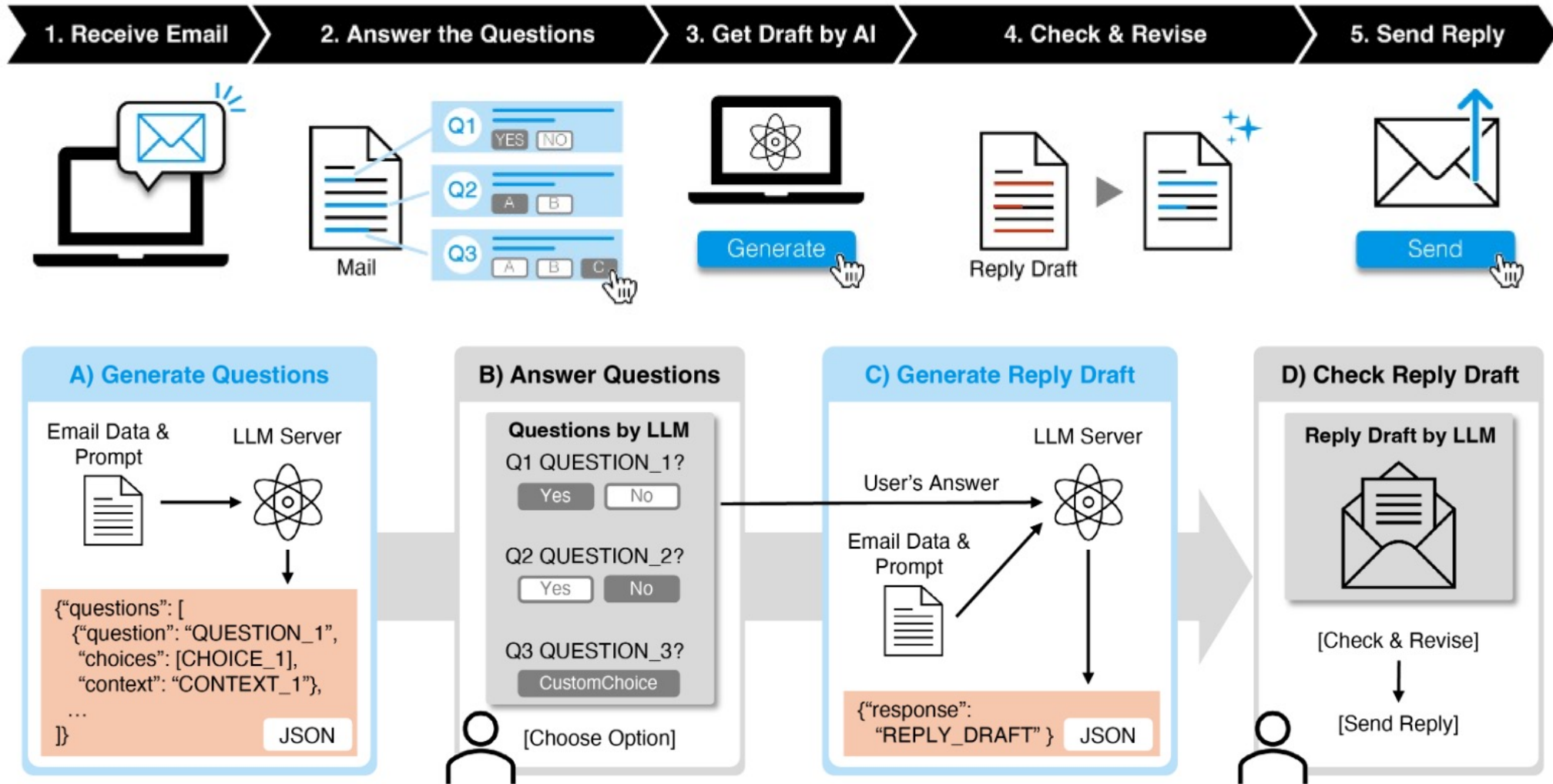
# Applications and Challenges

## Travel Planning (TravelPlanner+)



# Applications and Challenges

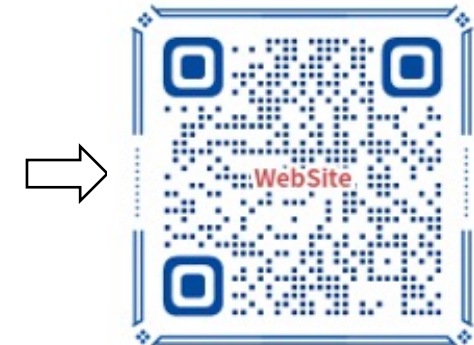
## □ Email Reply(ResQ)



# Tutorial Outline

- ⦿ Part 1: Introduction of WebAgents
- ⦿ Part 2: Preliminaries of AI Agents and LFM-based WebAgents
- ⦿ **Part 3: Architectures of WebAgents**
- **Coffee Break**
- **Part 4: Training of WebAgents**
- **Part 5: Trustworthy WebAgents**
- **Part 6: Future directions of WebAgents**

Website of this tutorial  
Check out the slides and more information!



# PART 3: Architectures of WebAgents

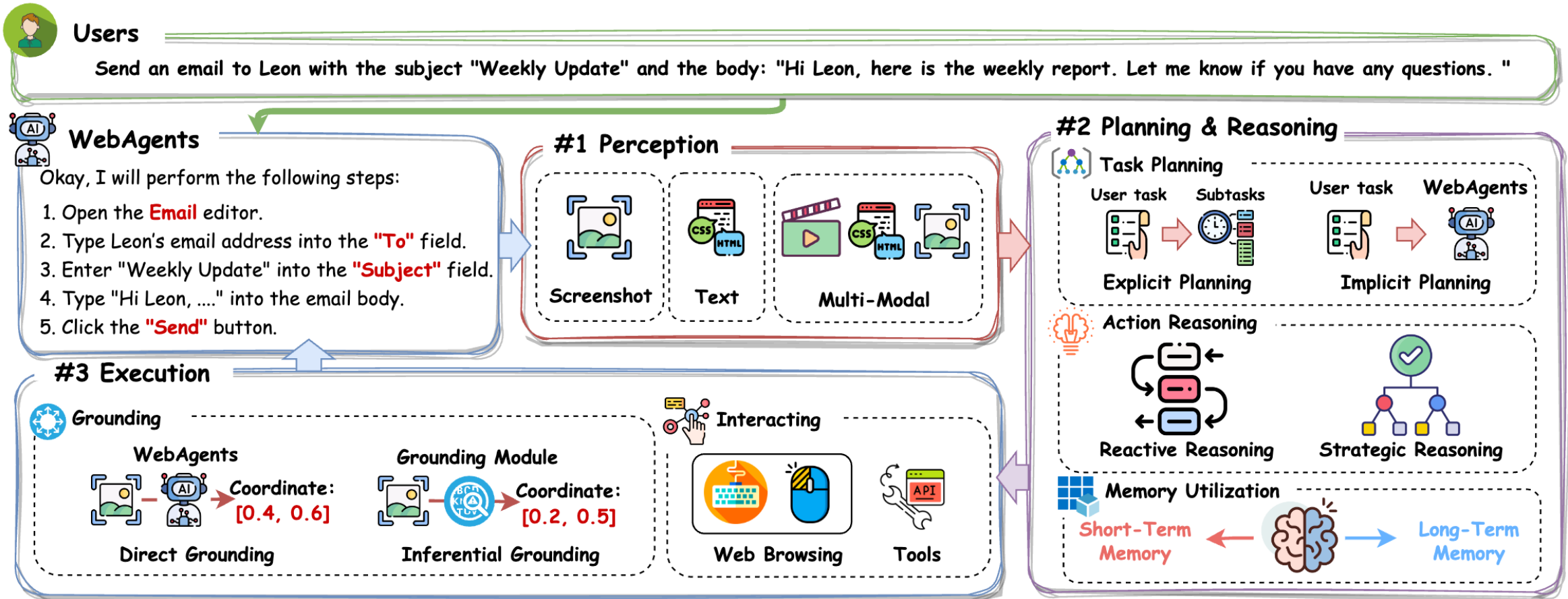


Website of this tutorial

- Perception
  - Text-based WebAgents
  - Screenshot-based WebAgents
  - Multi-modal WebAgents
- Planning & Reasoning
  - Task Planning
  - Action Reasoning
  - Memory Utilization
- Execution
  - Grounding
  - Interacting

# Architectures of WebAgents

- WebAgents, which contains three crucial processes: **Perception, Planning & Reasoning, and Execution.**



# PART 3: Architectures of WebAgents



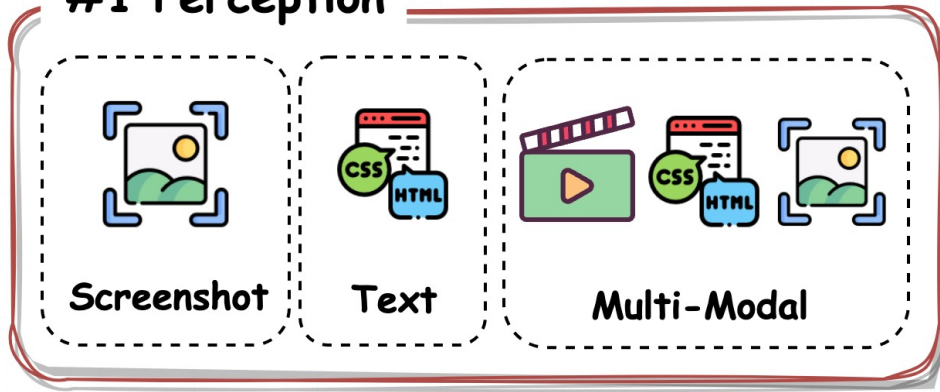
Website of this tutorial

- ⊙ **Perception**
  - **Text-based WebAgents**
  - **Screenshot-based WebAgents**
  - **Multi-modal WebAgents**
- **Planning & Reasoning**
  - Task Planning
  - Action Reasoning
  - Memory Utilization
- **Execution**
  - Grounding
  - Interacting

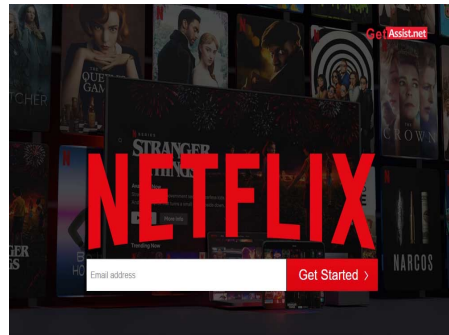
# Perception

- WebAgents are expected to accurately **perceive the external environment** and perform behavioral reasoning based on the dynamic environment.

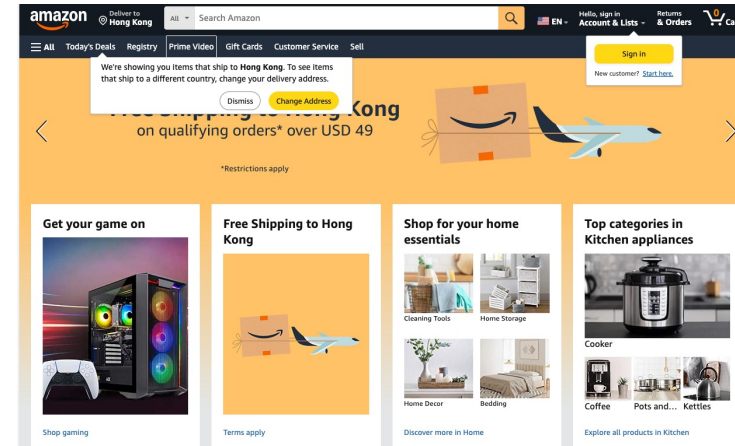
## #1 Perception



Video



## Screenshots



[1]: Delivering to Santa Clara 95050  
[2]:

...

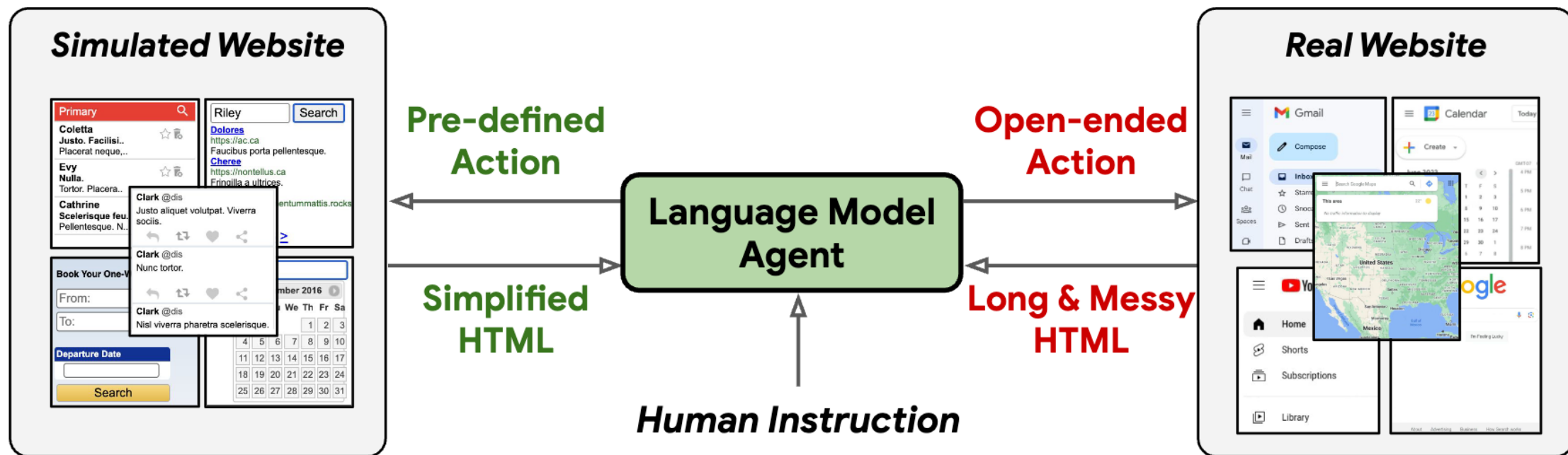
HTML

# Text-based WebAgents

## □ HTML-T5

Challenges in real-world web automation:

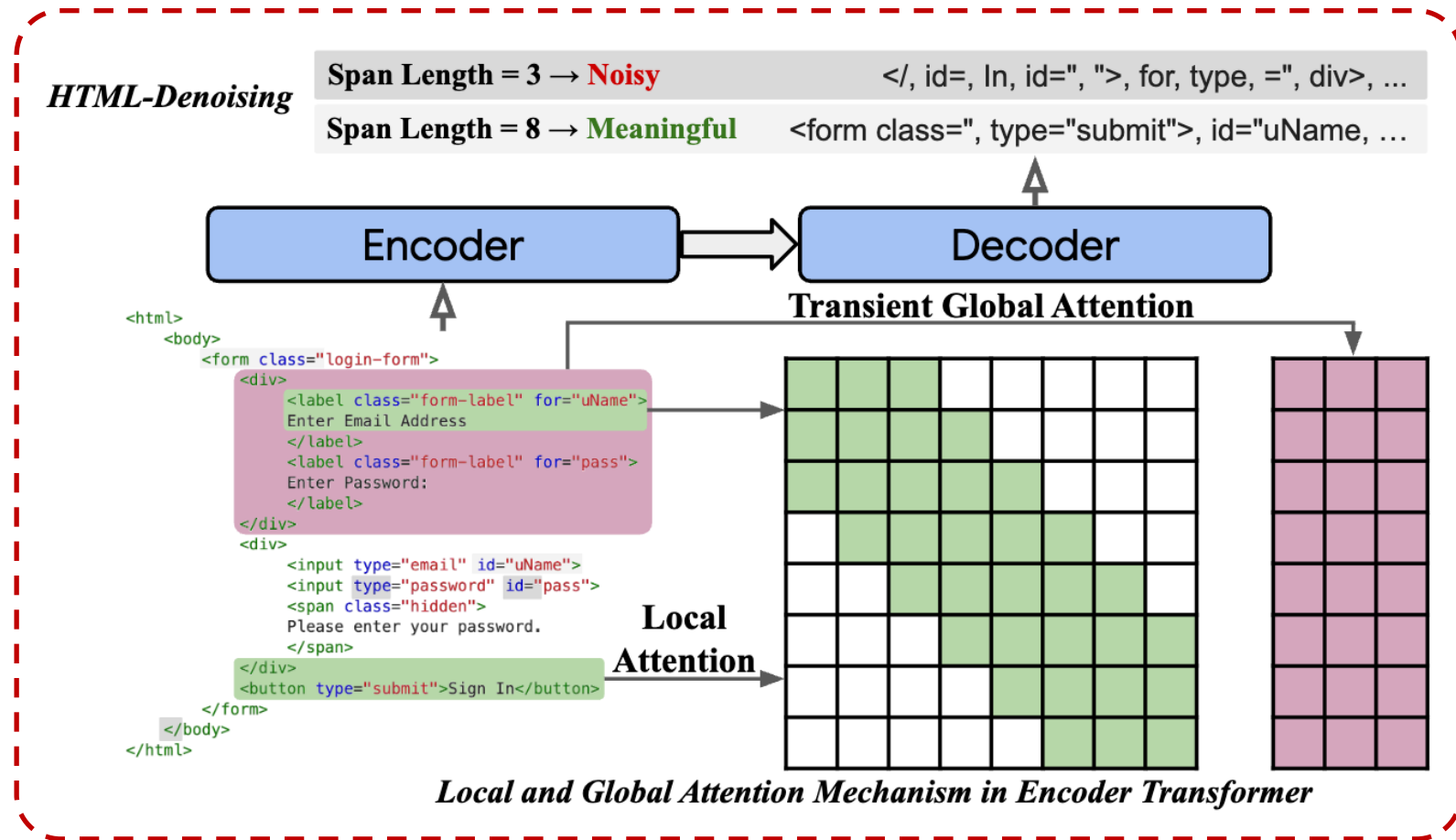
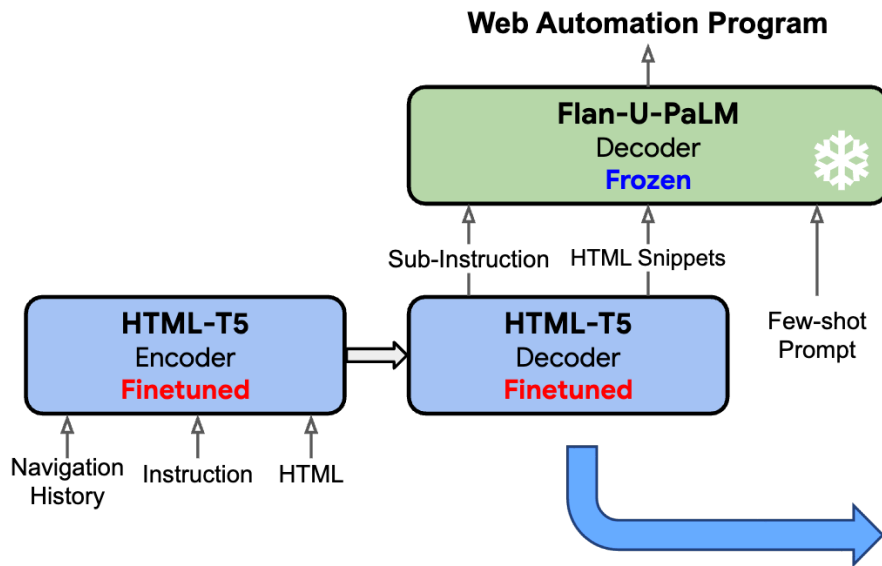
- Dynamic environments,
- open-ended actions,
- lengthy HTML documents,
- ...



# Text-based WebAgents

## HTML-T5

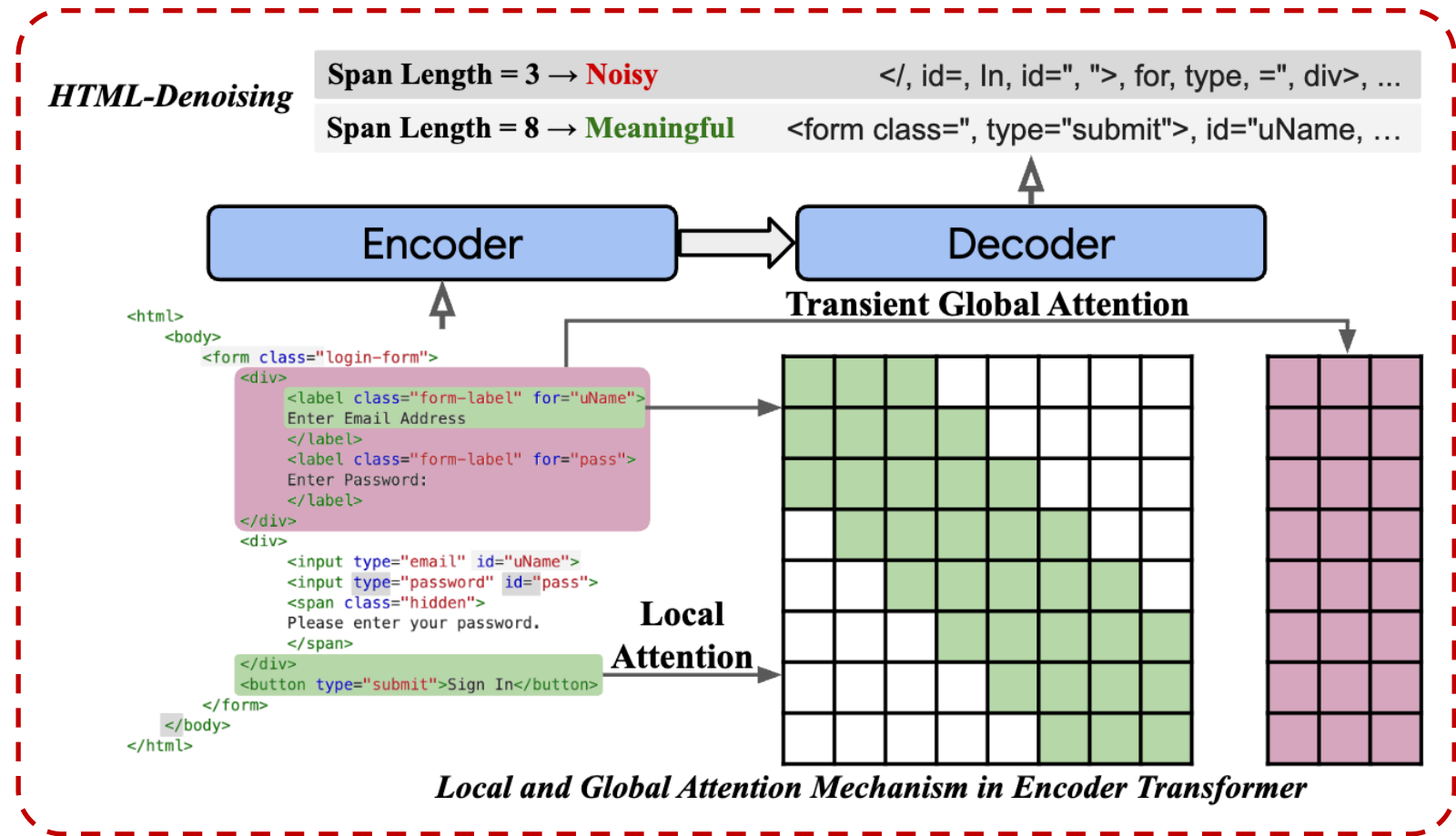
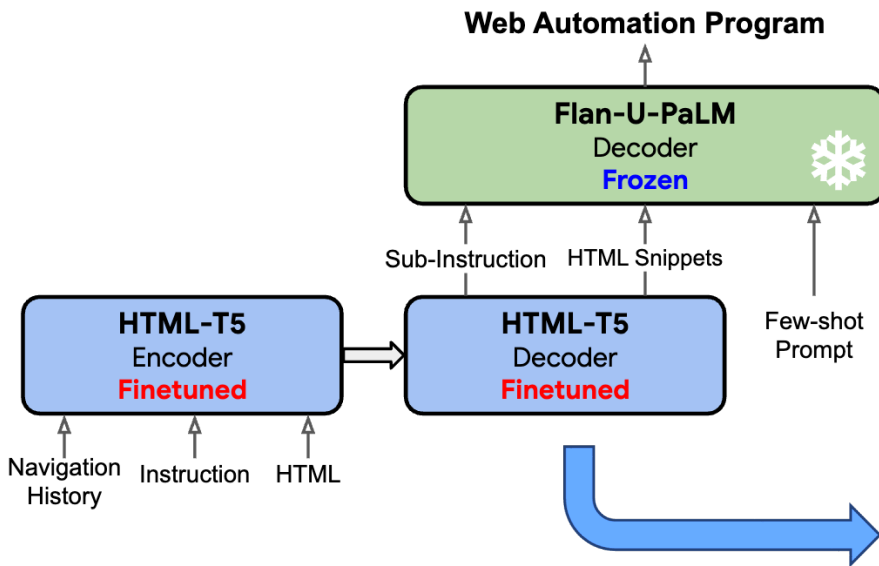
General-purpose LLMs do not fully utilize the HTML-specific information.



# Text-based WebAgents

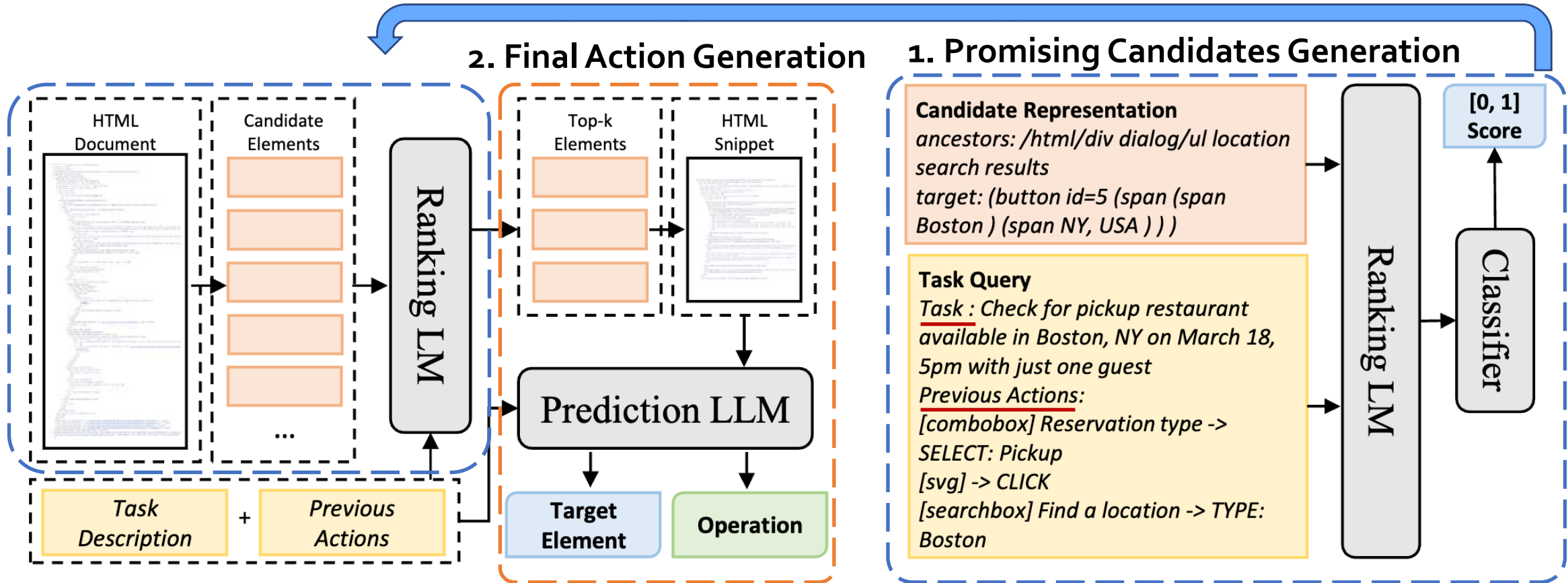
## HTML-T5

General-purpose LLMs do not fully utilize the HTML-specific information.



# Text-based WebAgents

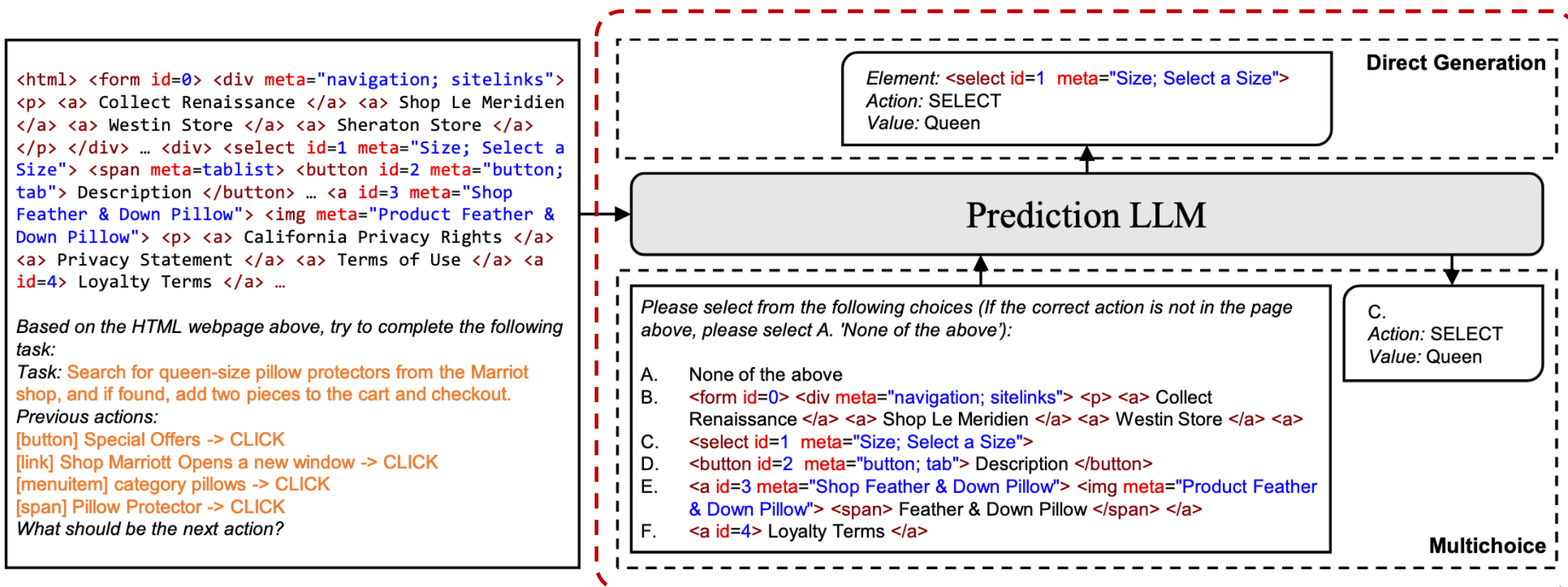
## □ Mind2Web



# Text-based WebAgents

## □ Mind2Web

Training LMs for discrimination rather than generation is more generalizable and sample-efficient for other grounding tasks.




# Screenshot-based WebAgents

- ❑ Textual representations usually vary across different environments and are **verbose**.
- ❑ To leverage the visual understanding capabilities of VLMs, numerous studies have integrated them into WebAgents, utilizing **screenshots** to perceive the environment.

## Text-based:

```
<form element_id="200">
  ...
  <label element_id="205">Last Name:</label>
  <input type="text" name="lastname" element_id="206">
  ...
  <input type="submit" value="Get Receipt" element_id="210">
  ...
</form>
```

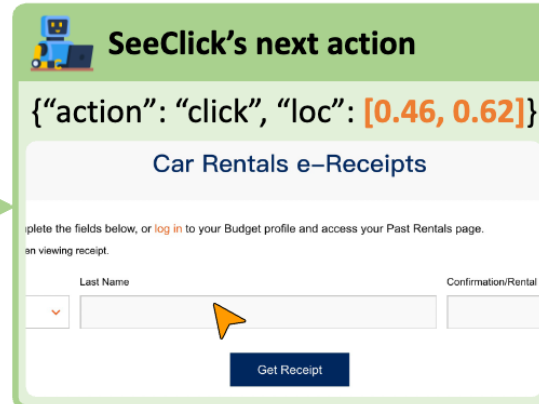
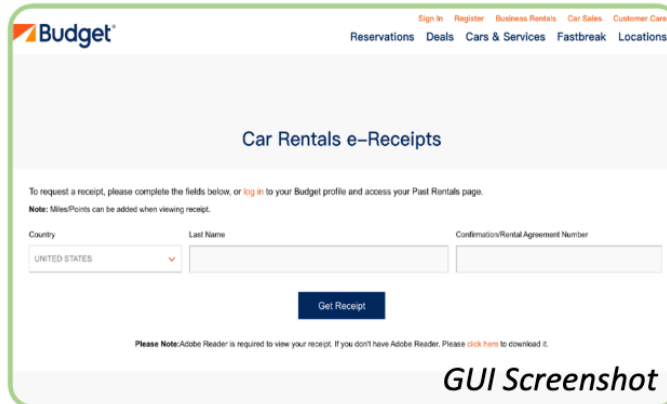
*Simplified HTML Code*

 **Text-based agent's next action**

Element: `<element_id=206>`  
Action: CLICK

```
# Selenium Code
element = driver.find_element(By.XPATH, '//*[@element_id="206"]')
element.click()
```

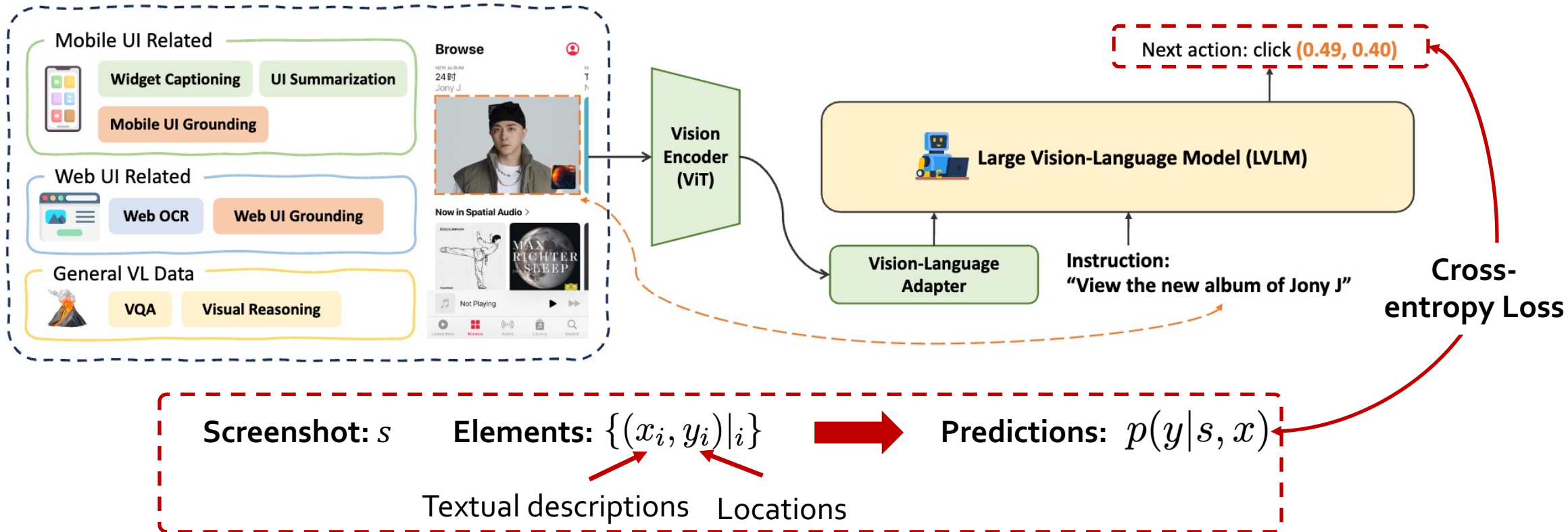
## Vision-based:



# Screenshot-based WebAgents

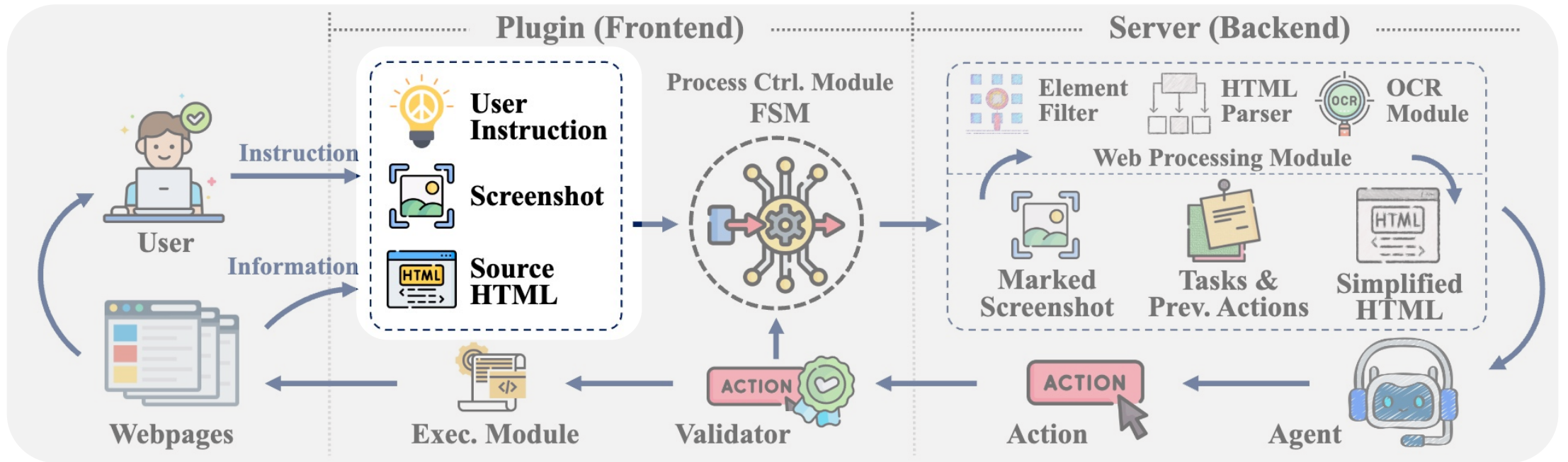
## SeeClick

**GUI grounding** - the capacity to accurately locate screen elements based on instructions, which is absent in current LVLMs.



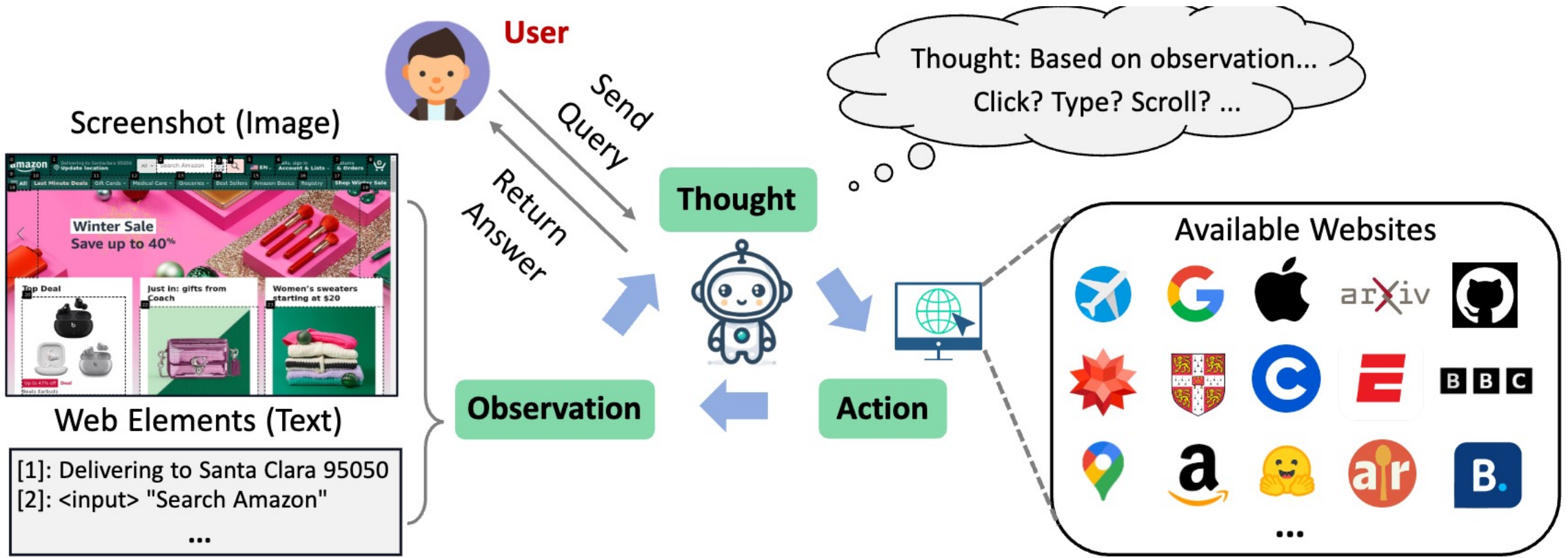
# Multi-Modal WebAgents

- ❑ **Leveraging multi-modal data**, combining their complementary strengths, can provide WebAgents with **a more comprehensive environmental perception**.



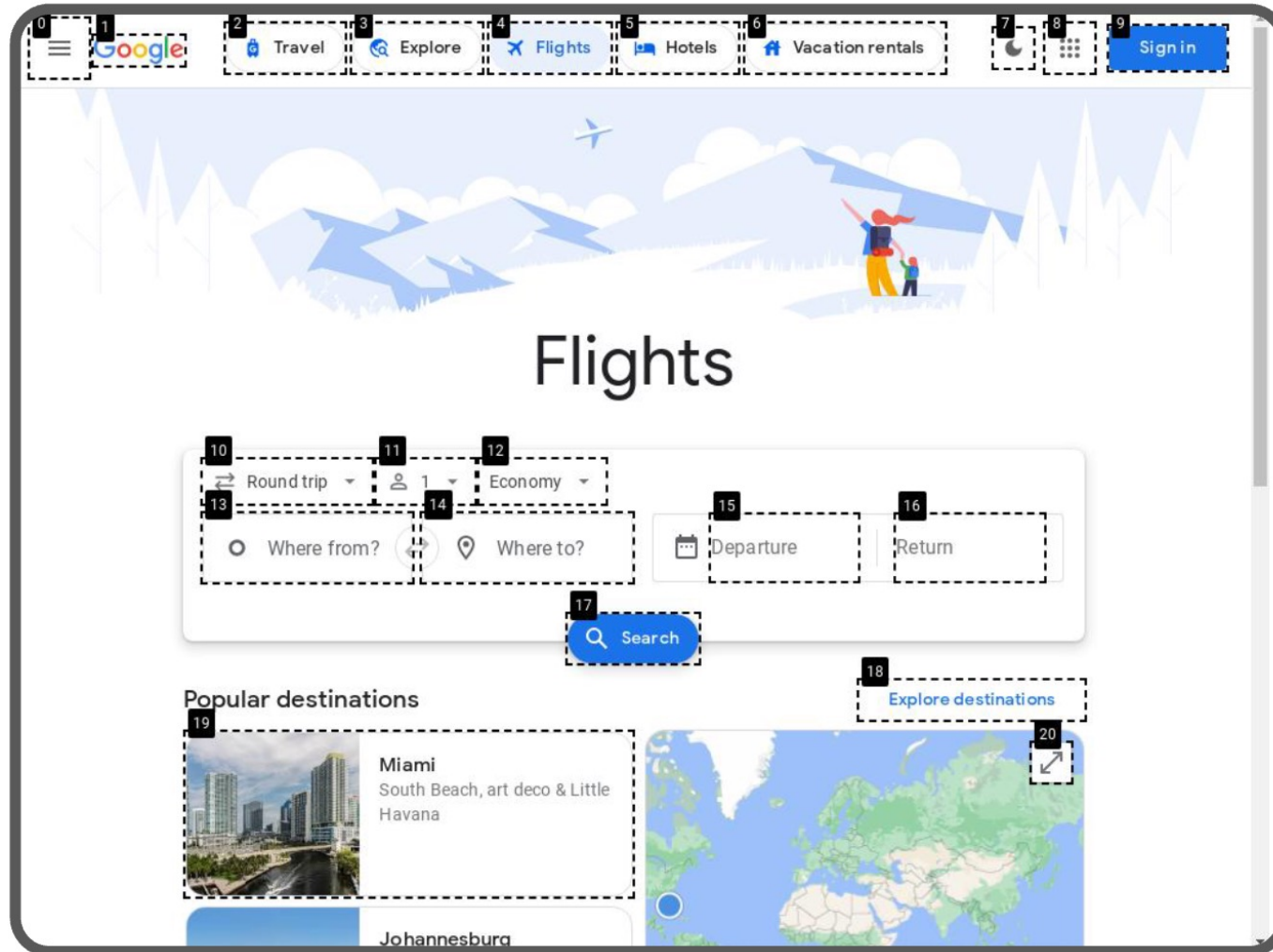
# Multi-Modal WebAgents

## WebVoyager



# Multi-Modal WebAgents

## WebVoyager



# Perception

## □ Text-based WebAgents

- Efficient
- ✓ ➤ Low resource
- Structured data
- ...

- Lengthy HTML documents,
- ✗ ➤ Noise
- ...

## □ Screenshot-based WebAgents

- Visual Context
- ✓ ➤ Align with human perception
- ...

- Resource intensive
- ✗ ➤ Additional vulnerability risk
- ...

## □ Multi-modal WebAgents

- Comprehensive information
- ✓ ➤ Robust
- ...

- High resource use
- ✗ ➤ Complex pipelines
- Potential redundancy
- ...

# Perception

## □ Text-based WebAgents

- Efficient
- ✓ ➤ Low resource
- Structured data
- ...

- Lengthy HTML documents,
- ✗ ➤ Noise
- ...

## □ Screenshot-based WebAgents

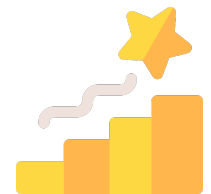
- Visual Context
- ✓ ➤ Align with human perception
- ...

- Resource intensive
- ✗ ➤ Additional vulnerability risk
- ...

## □ Multi-modal WebAgents

- Comprehensive information
- ✓ ➤ Robust
- ...

- High resource use
- ✗ ➤ Complex pipelines
- Potential redundancy
- ...



More Popular

# PART 3: Architectures of WebAgents



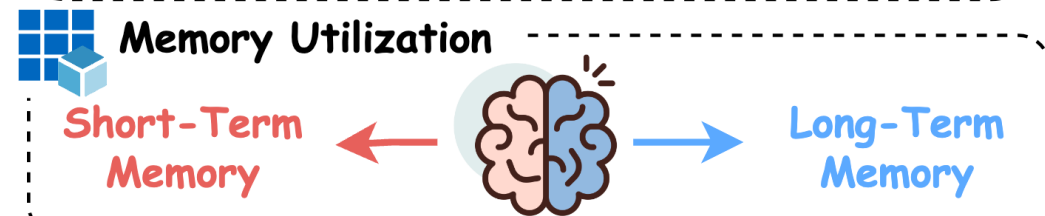
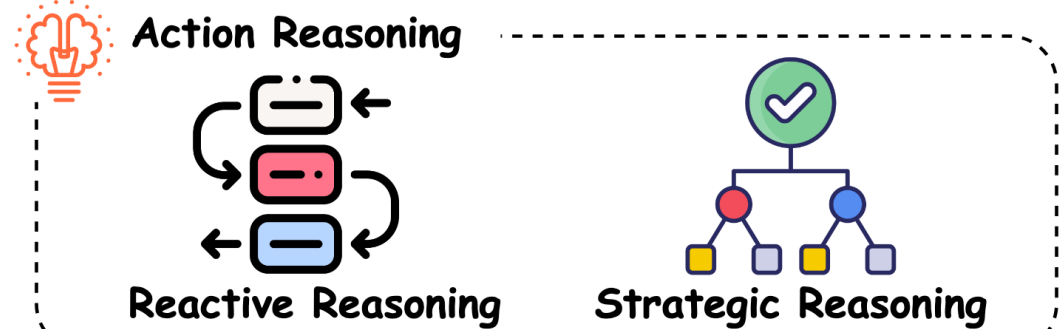
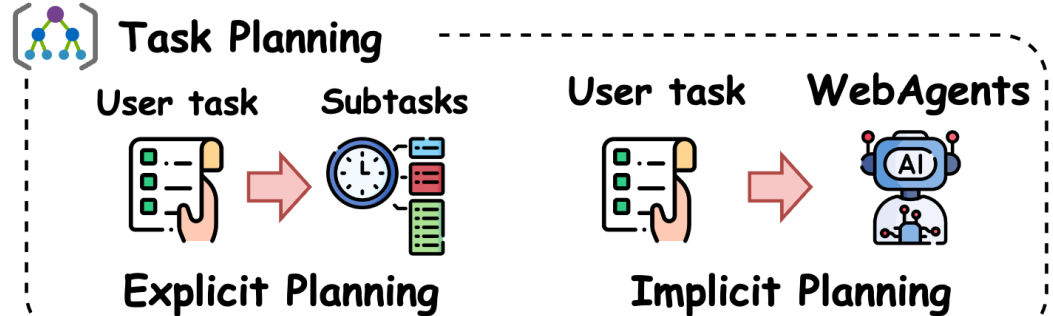
Website of this tutorial

- Perception
  - Text-based WebAgents
  - Screenshot-based WebAgents
  - Multi-modal WebAgents
- ⦿ **Planning & Reasoning**
  - **Task Planning**
  - **Action Reasoning**
  - **Memory Utilization**
- Execution
  - Grounding
  - Interacting

# Planning & Reasoning

- ❑ **1) Task Planning** reorganizes the user's instruction and sets sub-objectives.
- ❑ **2) Action Reasoning** guides WebAgents to generate appropriate actions.
- ❑ **3) Memory Utilization** equips WebAgents with internal or external information.

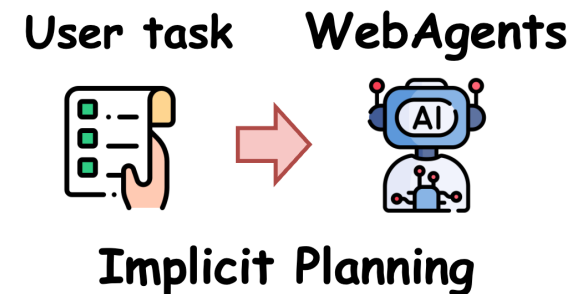
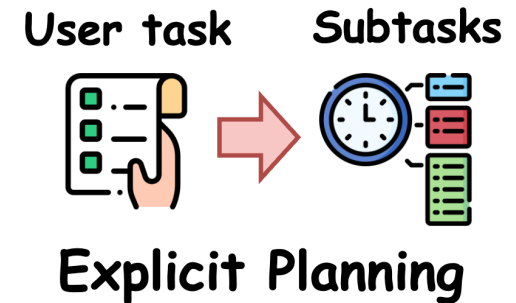
## #2 Planning & Reasoning



# Task Planning

**Task planning is to determine a sequence of steps that the agent should take to complete the user-defined task efficiently and effectively**

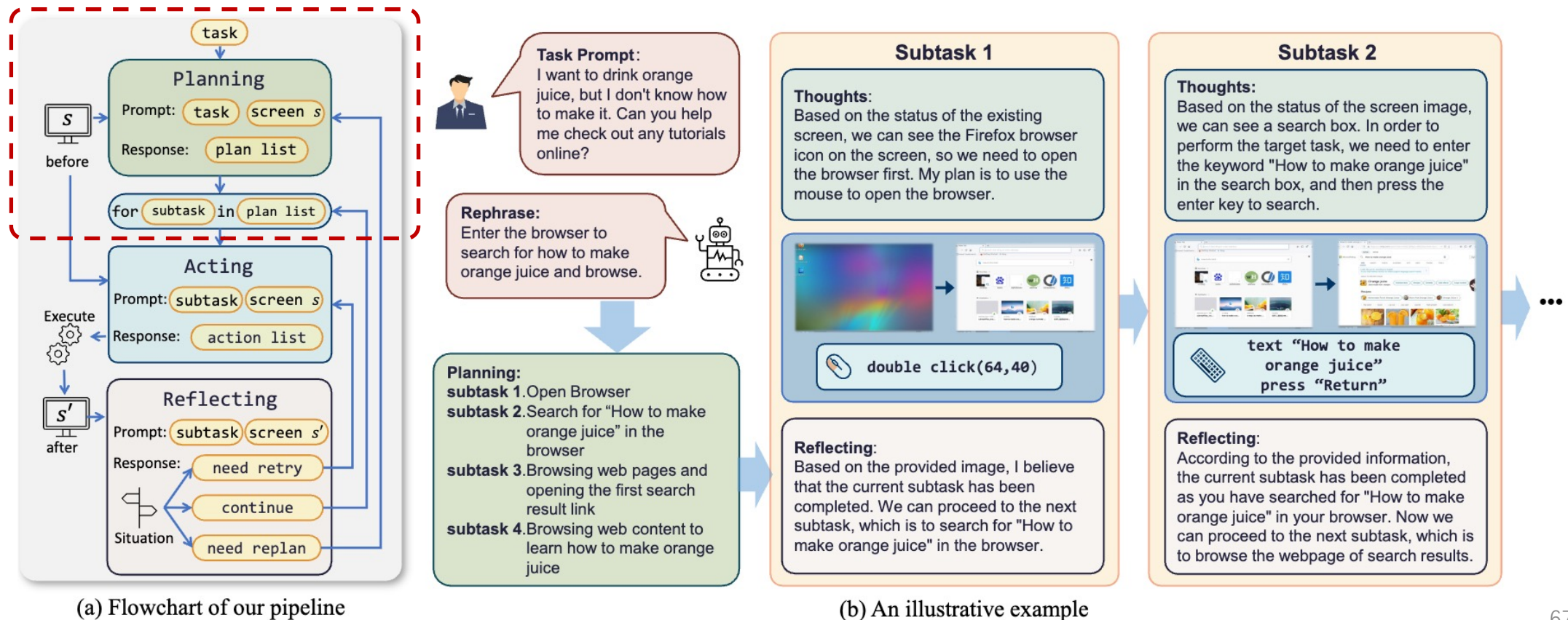
- ❑ **Explicit planning** methods usually decompose user instructions into **multiple sub-tasks**.
- ❑ **Implicit planning** does not contain an explicit task decomposition process.



# Task Planning – Explicit Planning

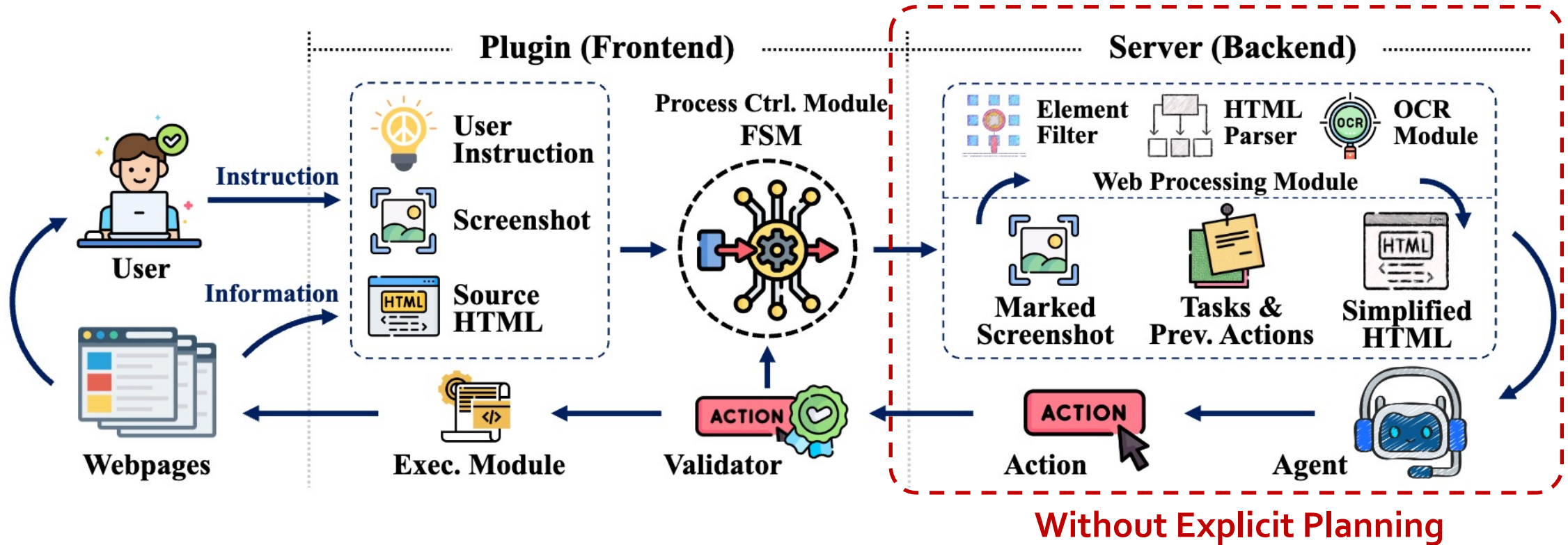
## □ ScreenAgent

Based on the current screenshot, the agent needs to **decompose the complex task relying on its own common-sense knowledge and computer knowledge.**



# Task Planning – Implicit Planning

## OpenWebAgent



# Planning

## □ Explicit Planning

- Controllable
- Precise
- Interpretable
- ...



- Inefficient
- Additional mechanisms, modules and processes
- ...



## □ Implicit Planning

- Efficient
- Easy to implement
- ...



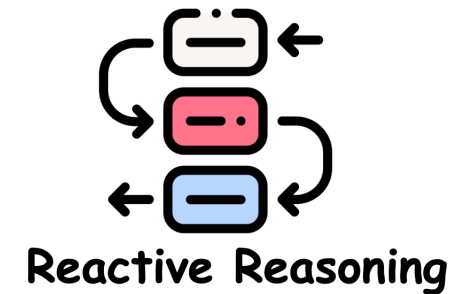
- Performance is highly dependent on the capabilities of LFM's
- Black-box process
- ...



# Action Reasoning

**Action reasoning** involves leveraging the agent's reasoning capabilities and current environmental observations to **infer the next action**.

- ❑ **Reactive reasoning:** WebAgents simply receive input prompts and directly generate the next actions **without additional operations**.

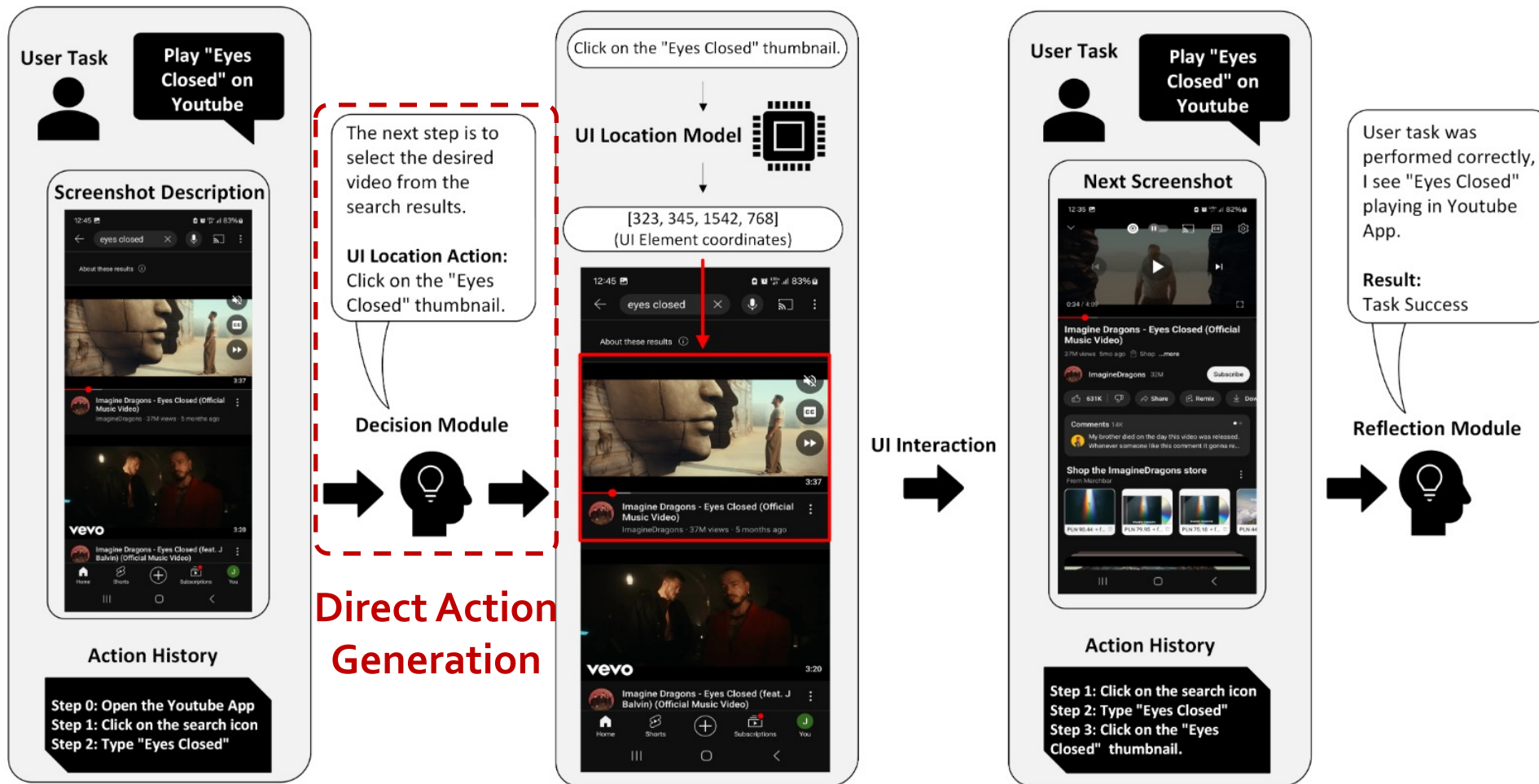


- ❑ **Strategic reasoning: Introduce additional operations** to enhance the agent's action reasoning capability.



# Action Reasoning – Reactive Reasoning

## ClickAgent

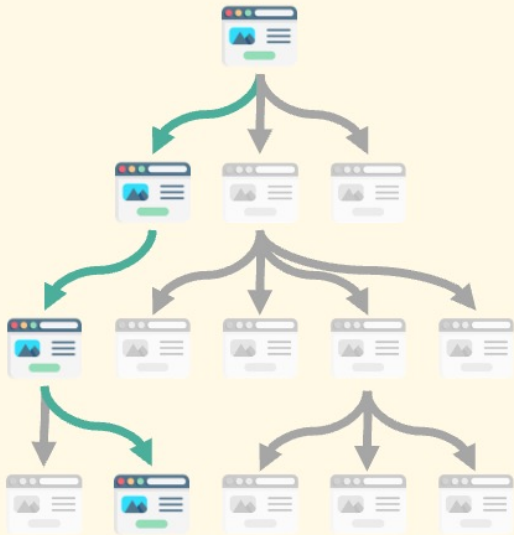


# Action Reasoning – Strategic Reasoning

## ❑ WebDremer

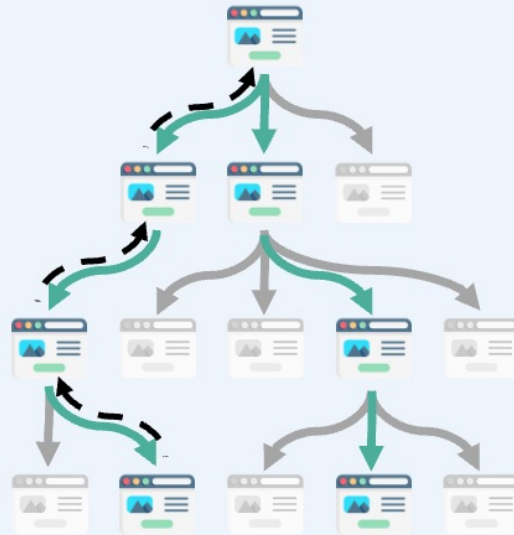
**LLM-driven exploration strategy** to simulate and predict the outcomes of candidate actions before execution using natural language descriptions.

(a) reactive



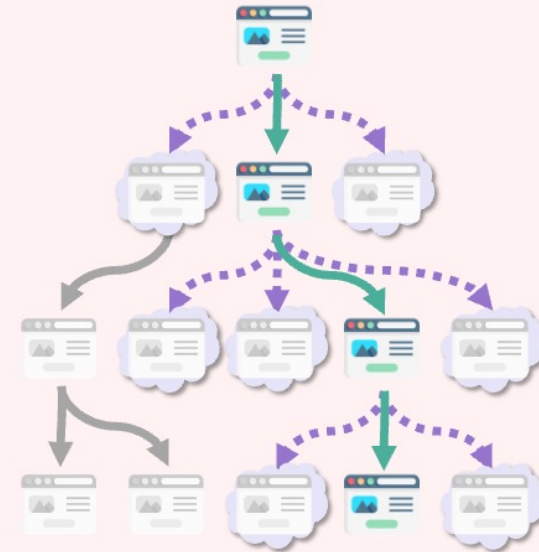
Locally optimal actions ❌

(b) tree search with real interactions



Backtracking is infeasible ❌

(c) model-based planning

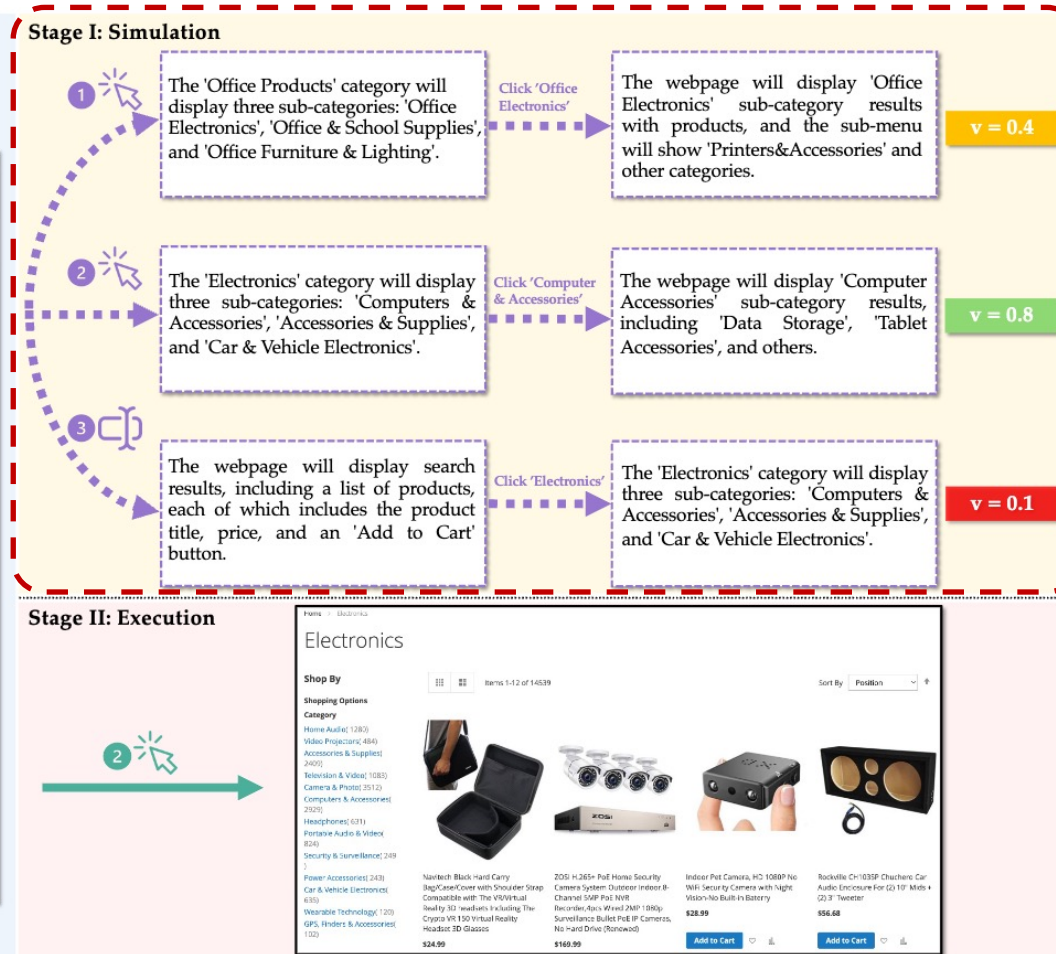
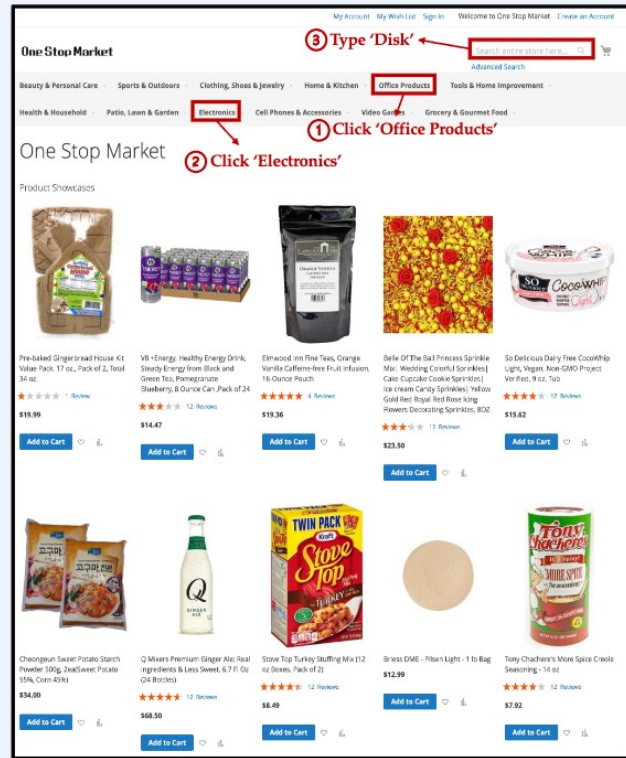


Backtracking is infeasible ✅

# Action Reasoning – Strategic Reasoning

## WebDreamer

Please navigate to the 'Data Storage' category and purchase the least expensive disk with 512GB of storage.



### Algorithm 1: WEBDREAMER

**Input:** Instruction  $I$ ; initial observation  $o_0$

**Output:** Sequence of actions  $a_0, a_1, \dots, a_T$   
 $t \leftarrow 0$ ;

**while True do**

$\mathcal{A}_t \leftarrow \text{get\_candidate}(I, o_t)$ ;

$\mathcal{A}'_t \leftarrow \text{self\_refine}(\mathcal{A}_t)$ ;

$a_t = \arg \max_{a \in \mathcal{A}'_t} \text{score}(\text{sim}(o_t, a))$ ;

$o_{t+1} \leftarrow \text{execute}(a_t)$ ;

$t \leftarrow t + 1$ ;

**if**  $\text{termination\_check}() = \text{True}$  **then**  
 | **break**;

**end**

**end**

**Return** result;

# Reasoning

## □ Reactive Planning

- Efficient
- ✔ Easy to implement
- ...

- ✘ Performance is highly dependent on the capabilities of LFM
- ...

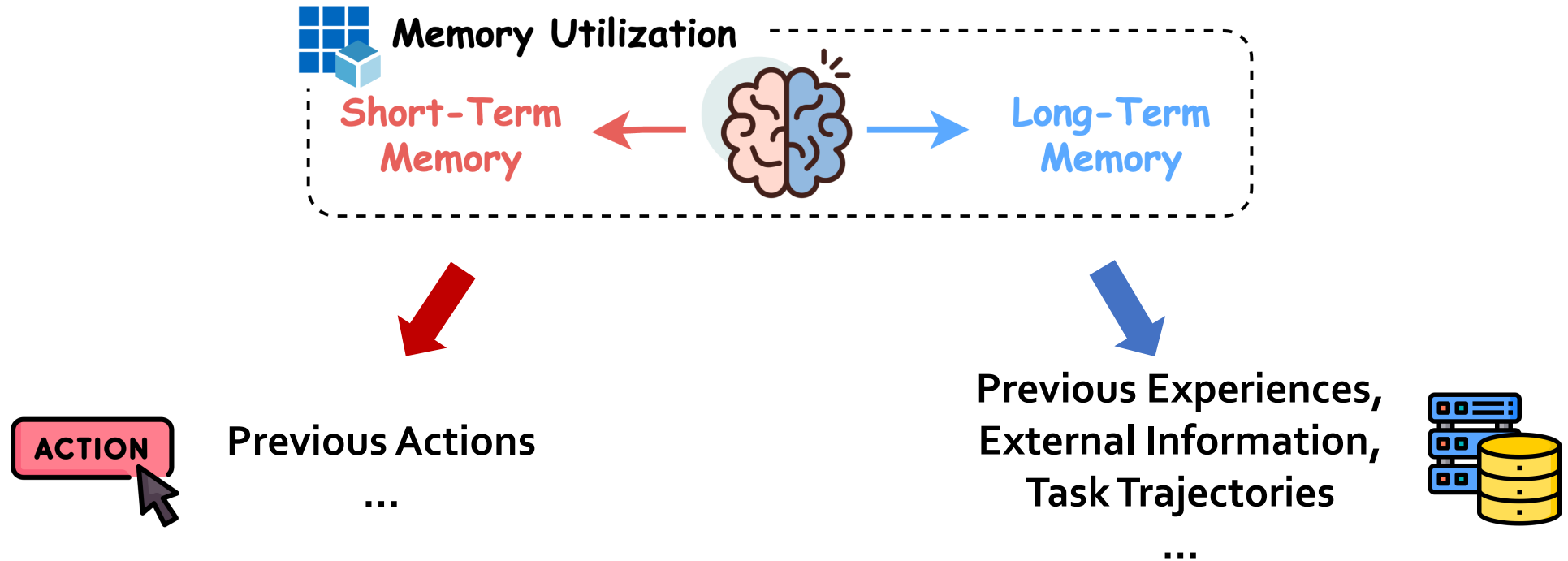
## □ Strategic Reasoning

- More accurate action prediction
- ✔ ...

- ✘ Additional computational cost
- Design sophisticated mechanisms.
- ...

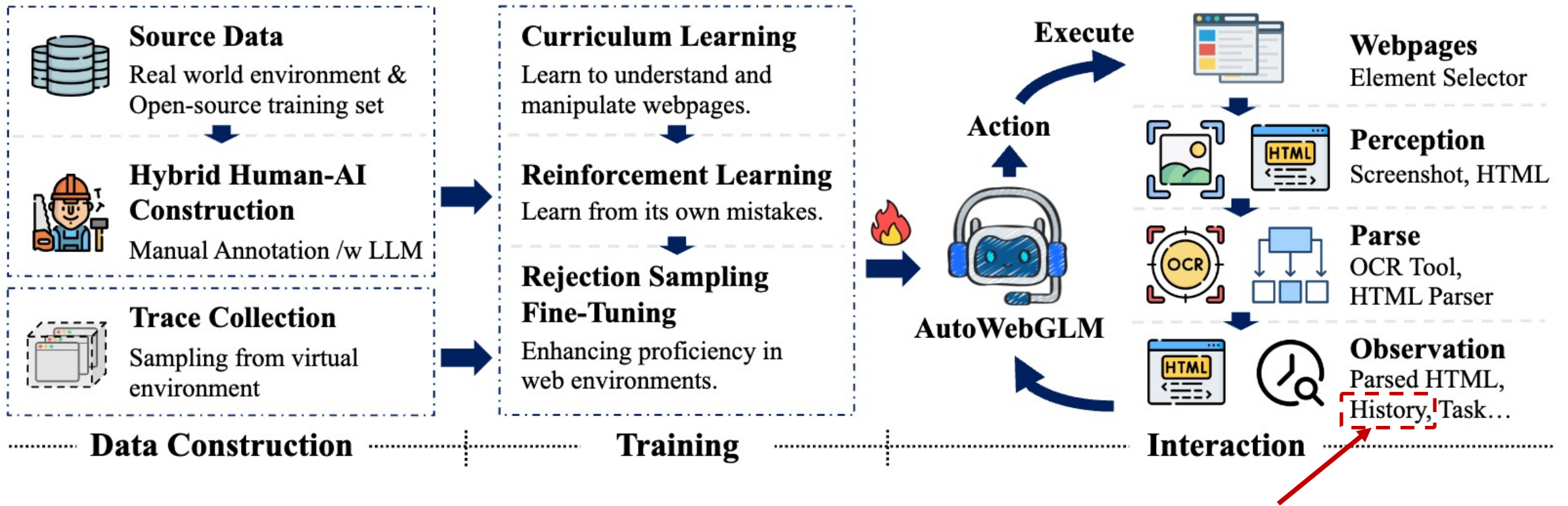
# Memory Utilization

- Beyond task planning and action reasoning, **the effective utilization of memory is another key factor contributing to the powerful capabilities of WebAgents.**



# Memory Utilization – Short-Term

## AutoWebGLM

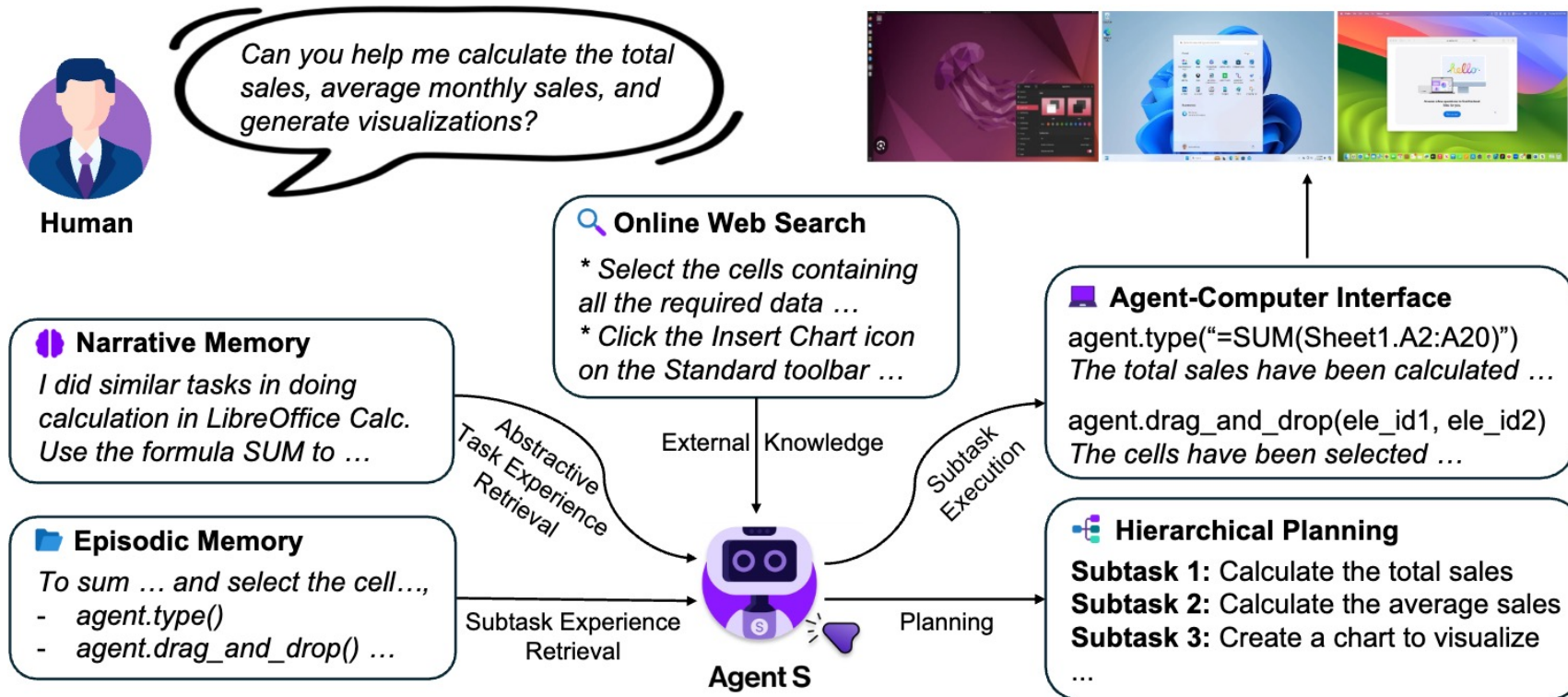


The best solution to inform the agent of past operations is explicitly providing it.

# Memory Utilization – Long-Term

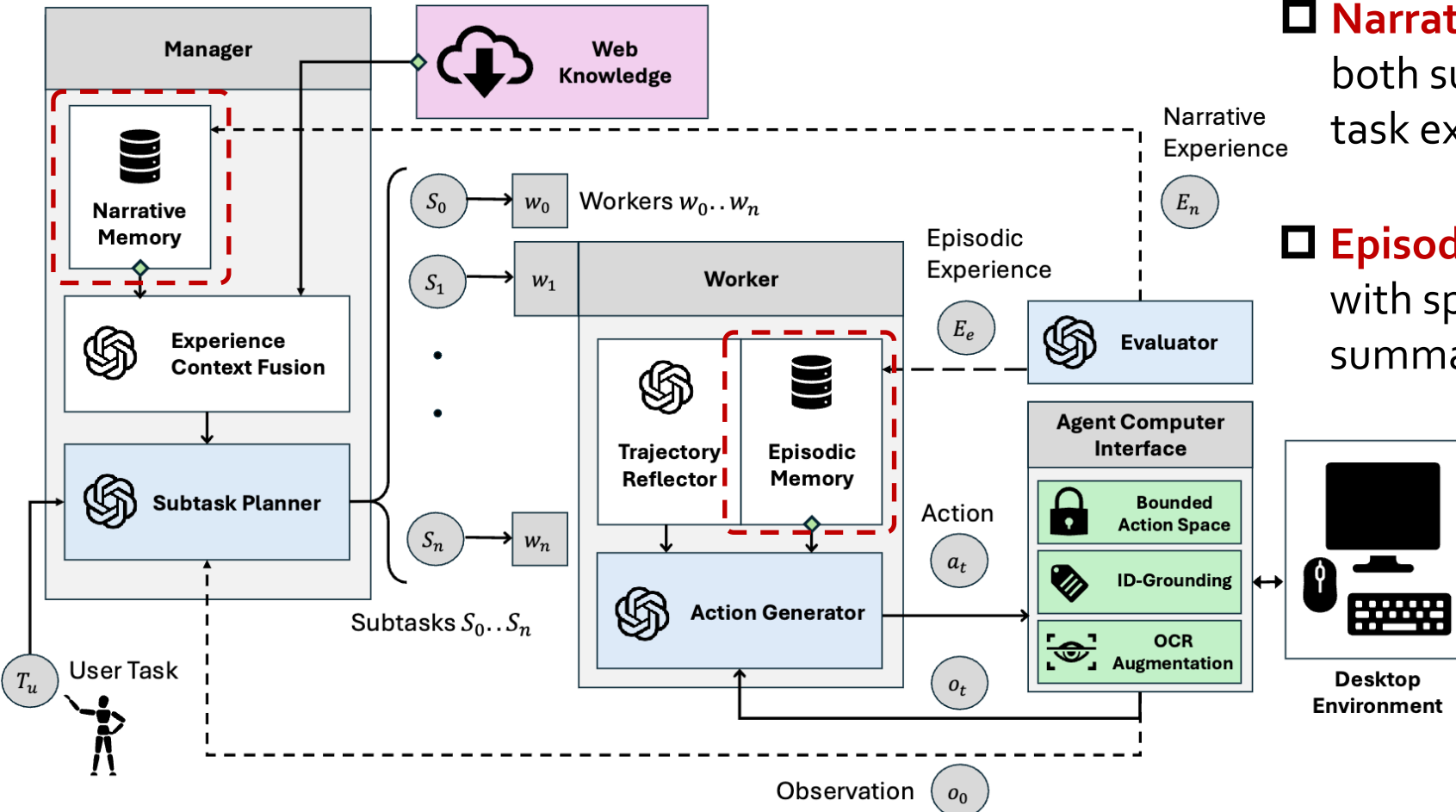
## AGENT S

- Leverages **Online Web Knowledge** and past experiences stored in **Narrative Memory** to decompose the complex, long-horizon task into a structured plan of manageable subtasks
- Retrieves step-by-step subtask experience from **Episodic Memory** to refine the actions.



# Memory Utilization – Long-Term

## AGENT S

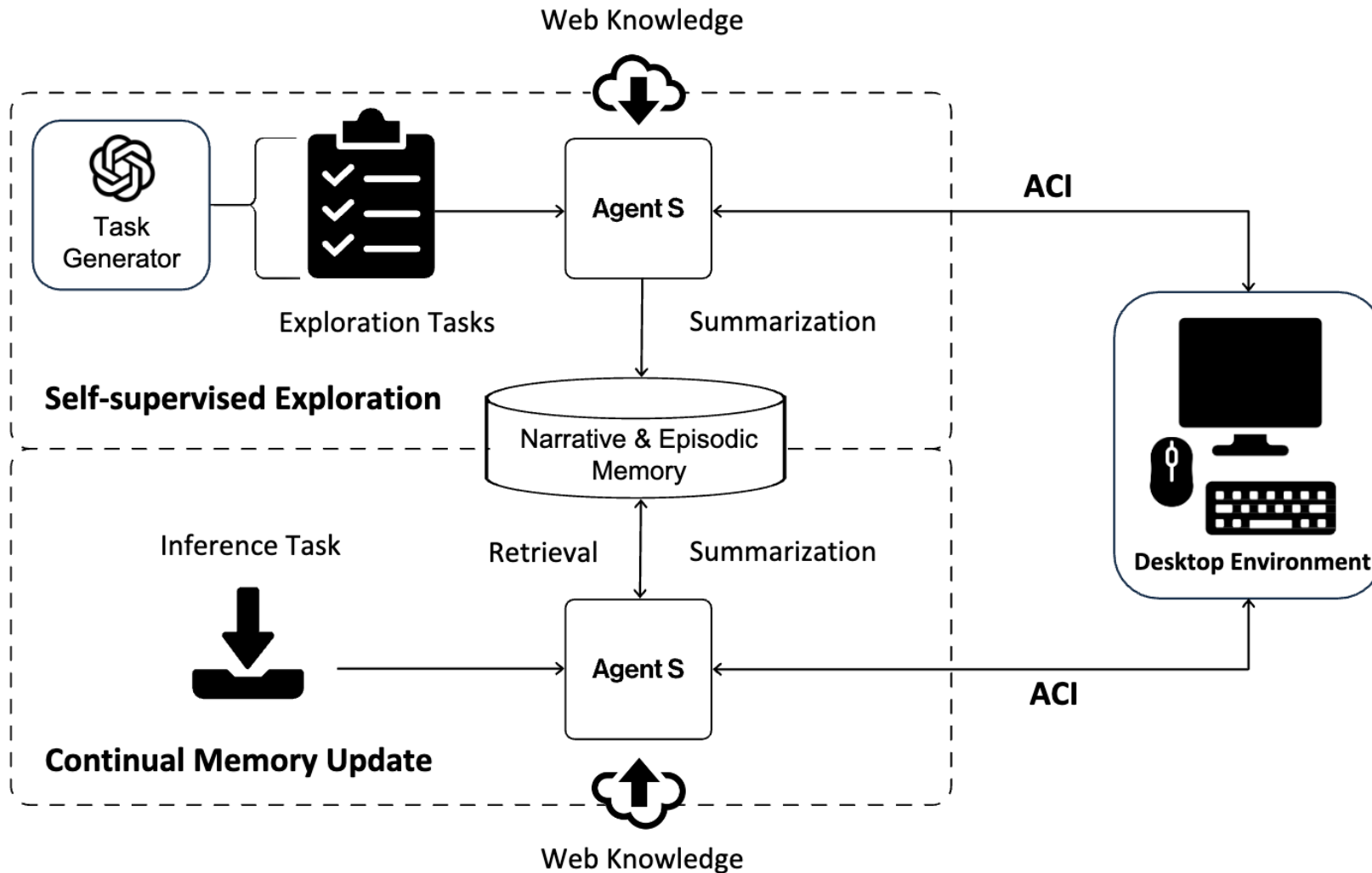


□ **Narrative Memory:** includes summaries of both successful and failed trajectories as task experiences.

□ **Episodic Memory:** includes a complete plan with specific grounding actions and only summaries from the subtask trajectories

# Memory Utilization – Long-Term

## AGENT S



# Memory Utilization

## ☐ Short-term Memory

- Efficient
- ✔ ➤ Short context
- ...

- Limited information
- ✘ ➤ Fixed knowledge
- ...

## ☐ Long-term Memory

- More accurate task completion
- ✔ ➤ External and up-to-date knowledge
- ...

- Lengthy context
- ✘ ➤ Sophisticated mechanisms to store and leverage the memory
- ...

# PART 3: Architectures of WebAgents

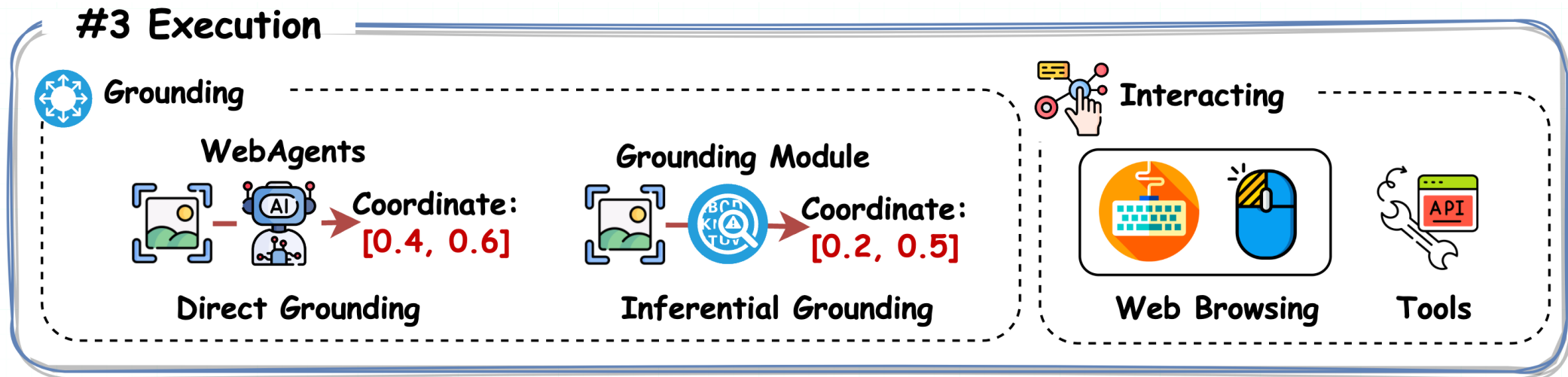


Website of this tutorial

- Perception
  - Text-based WebAgents
  - Screenshot-based WebAgents
  - Multi-modal WebAgents
- Planning & Reasoning
  - Task Planning
  - Action Reasoning
  - Memory Utilization
- **Execution**
  - **Grounding**
  - **Interacting**

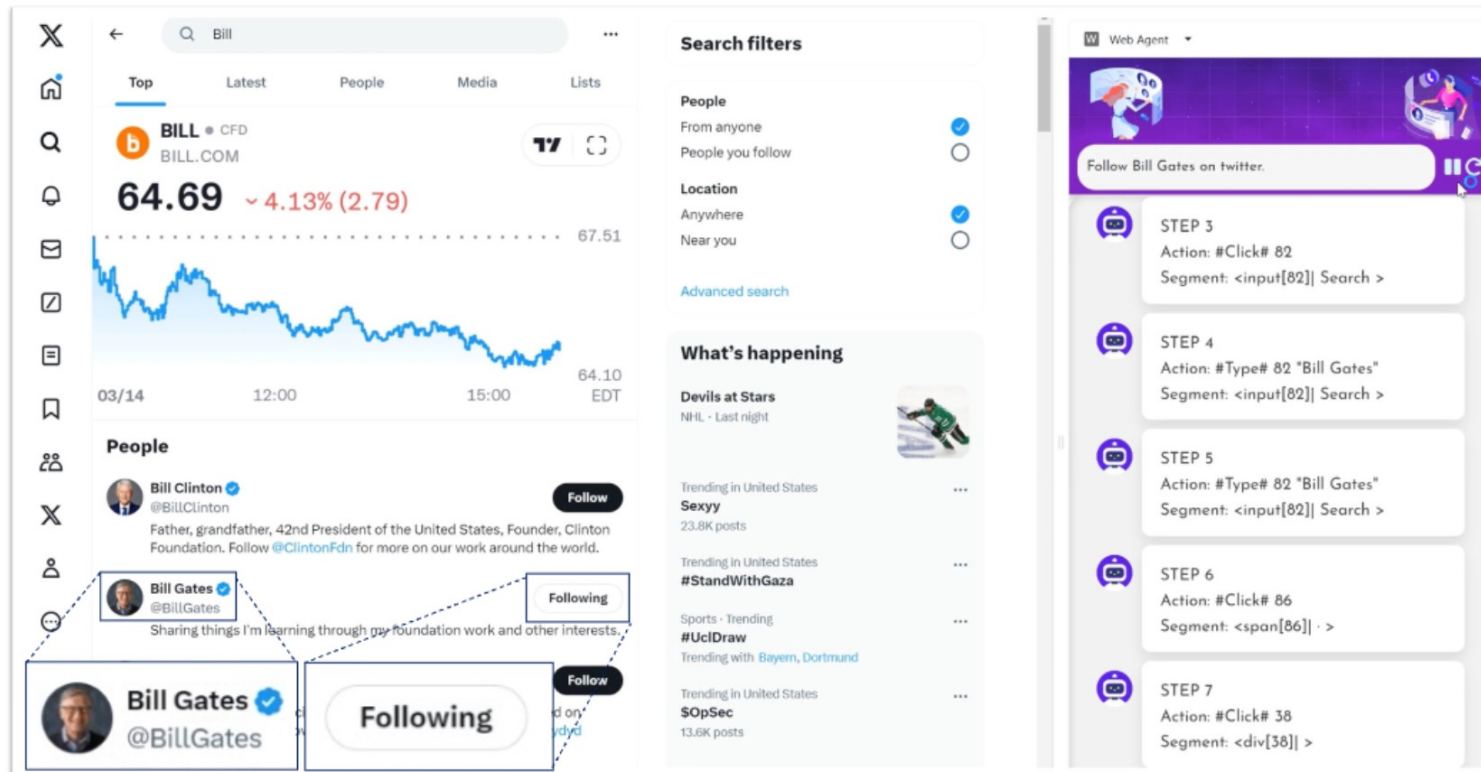
# Execution

- ❑ The final step for WebAgents to complete the user's command is to **interact with the webpages and execute the generated actions.**

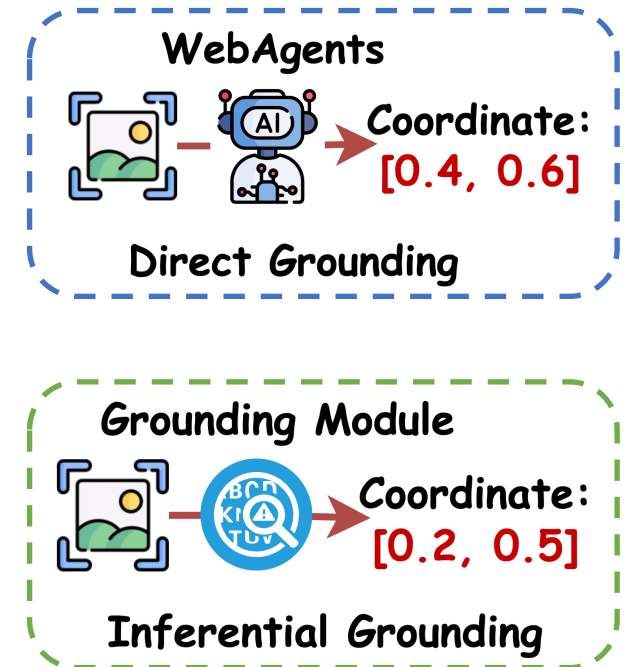


# Grounding

- Since webpages often contain numerous interactive elements, **selecting the correct element to execute the generated action** is crucial for completing the user's task.

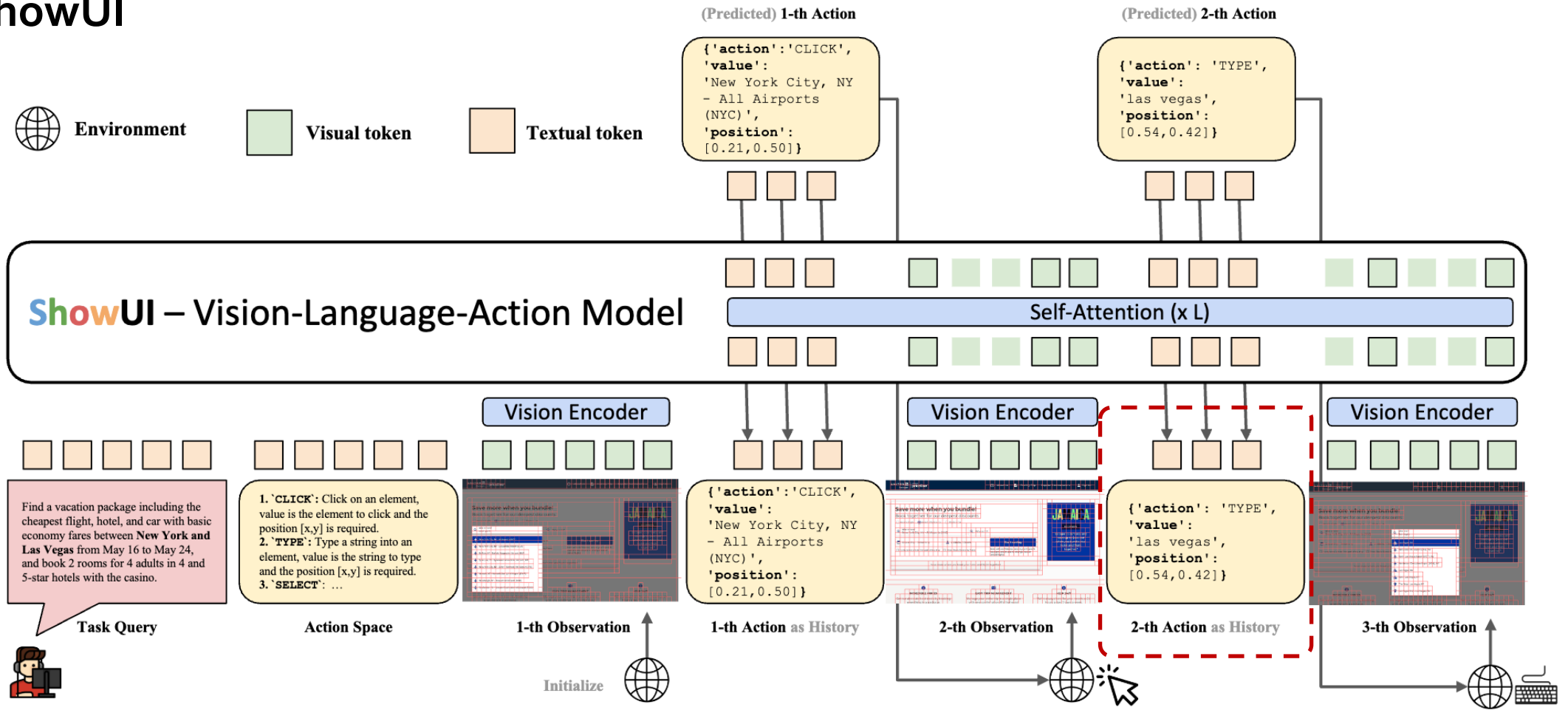


(a) Follow Bill Gates on X (Twitter).



# Grounding – Direct Grounding

## ShowUI

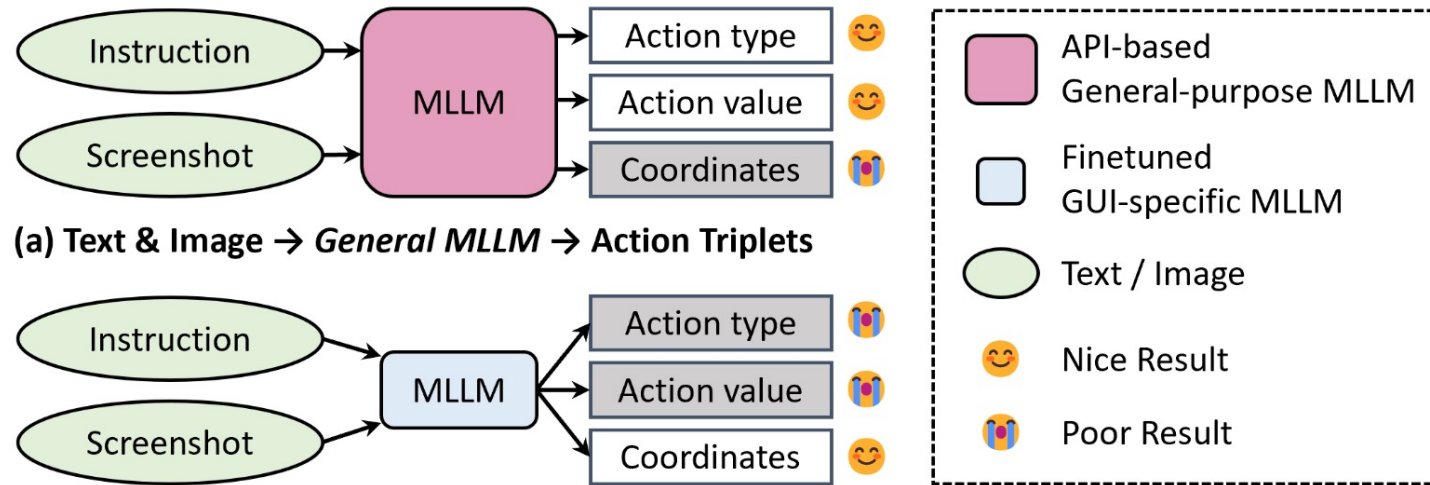


Contains both generated actions and the target element position

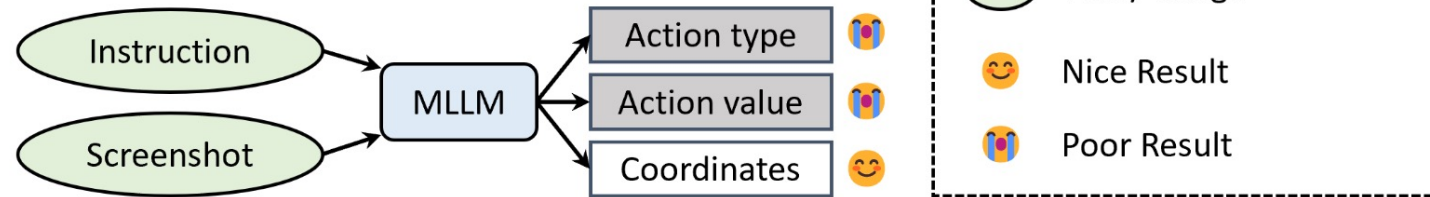
# Grounding – Inferential Grounding

## ❑ Ponder & Press

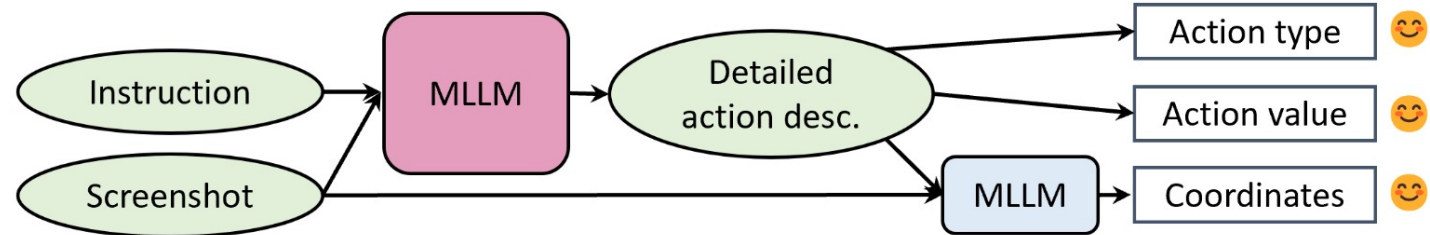
User instructions are directly mapped to action triplets in a single inference, which is **difficult due to the significant difference between the textual nature of actions and values, and the numerical nature of pixel coordinates.**



(a) Text & Image → General MLLM → Action Triplets



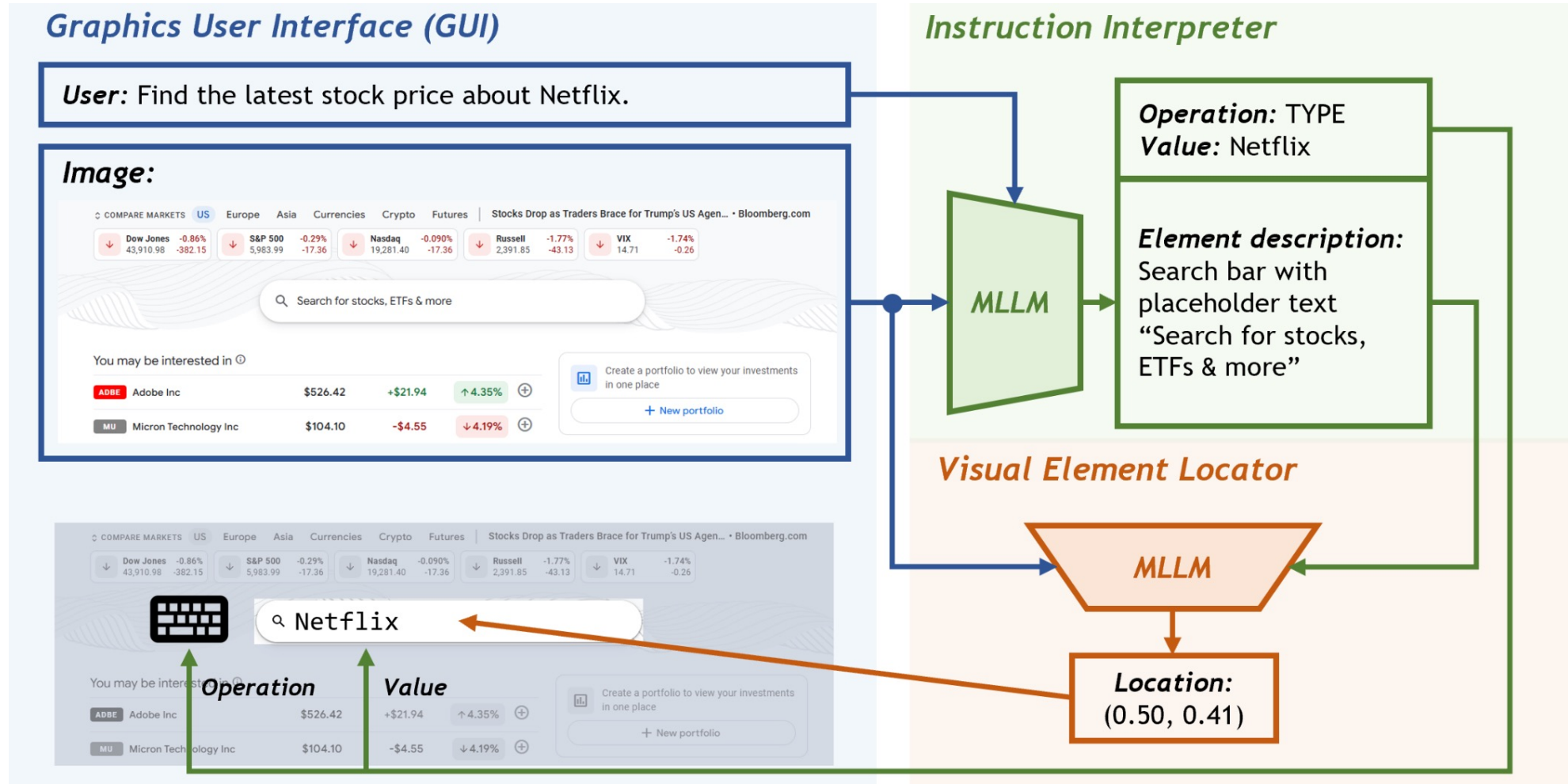
(b) Text & Image → GUI-Spec. MLLM → Action Triplets



(c) Text & Image → General MLLM → Action Description → GUI-Spec. MLLM → Action Triplets

# Grounding – Inferential Grounding

## ❑ Ponder & Press



# Grounding

## □ Direct Grounding

- ✔ Efficient
- ✔ Easy to implement
- ✔ ...

- ✘ Grounding accuracy is dependent on the capabilities of LFM
- ✔ ...

## □ Inferential Grounding

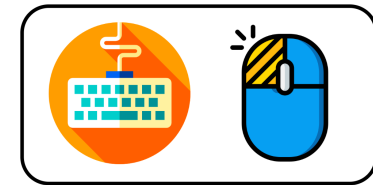
- ✔ More accurate element location
- ✔ ...

- ✘ Sophisticated mechanisms to locate the correct element
- ✔ ...

# Interacting

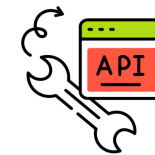
WebAgents need to **interact with the target element using the generated actions.**

- ❑ Web browsing-based methods utilize **typical actions that humans employ when navigating websites.**



Web Browsing

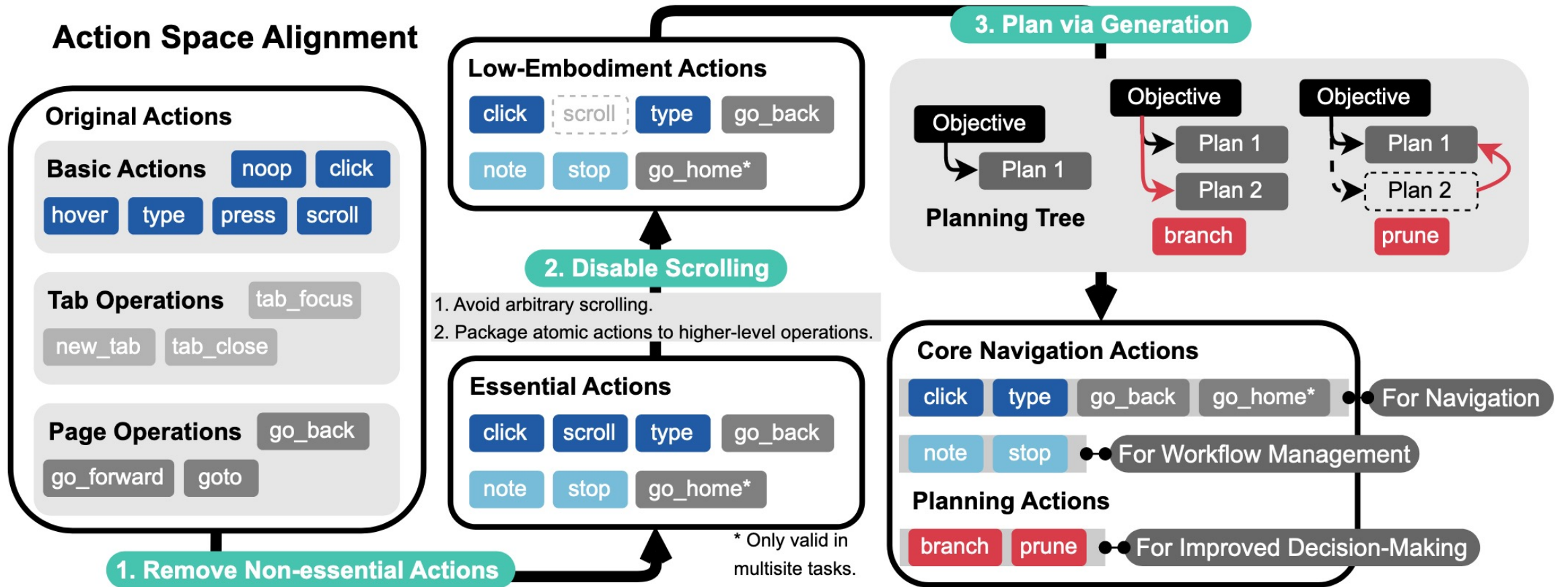
- ❑ Tool-based methods involve using **additional tools, such as APIs,** to interact with the webpages.



Tools

# Interacting – Web Browsing

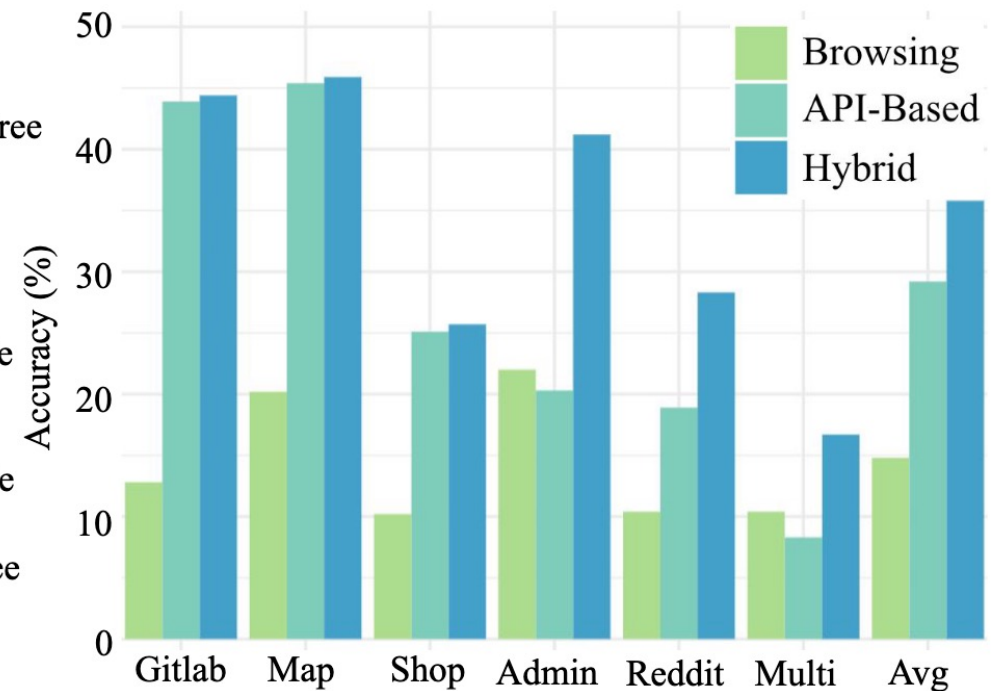
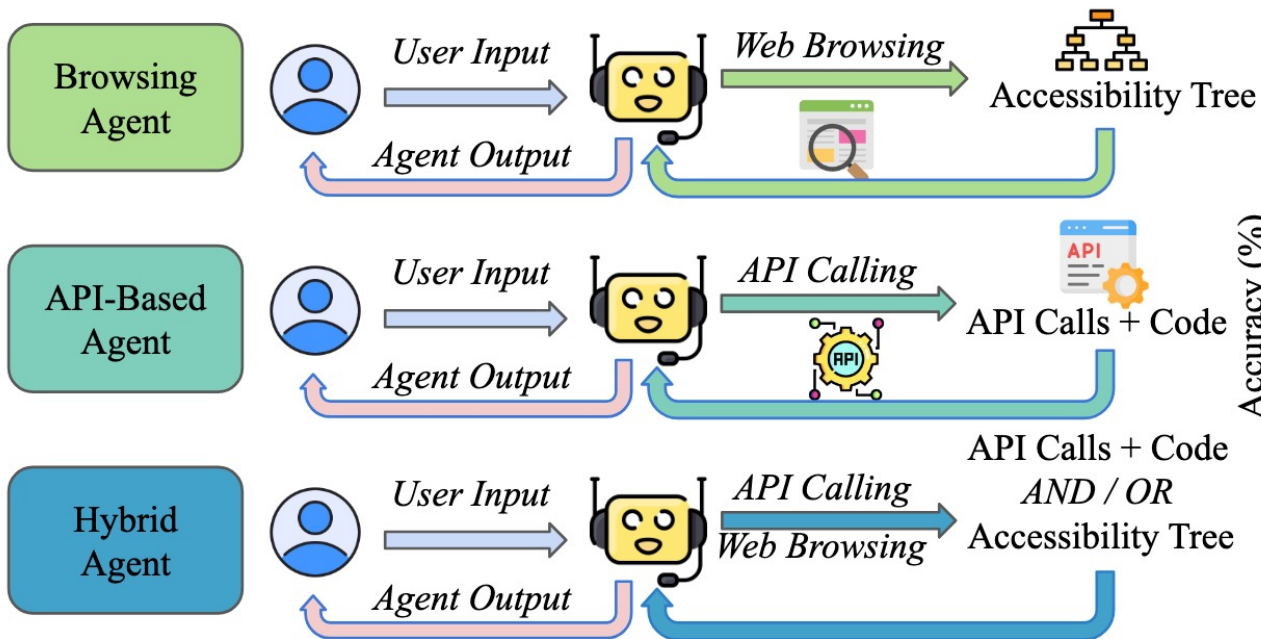
## AGENTOCCAM



# Interacting – Tool

## □ API-Calling Agents

Regardless of the interaction method with websites, there is no getting around the fact that **these sites were originally designed for humans, and may not be the ideal interface for machines.**



# Interacting – Tool

## □ API-Calling Agents



API Documentation	<pre># Commits ## GET /api/{id}/commits: Get a list of commits in a project.   Attribute   Type   Description     `id`   integer/string   The ID or path of the project.     `since`   string   Only commits after or on this date.     `until`   string   Only commits before or on this date.   Output: JSON containing all commits that meet the given criteria.</pre>
API Calling	<pre>&lt;execute_ipython&gt; requests.get('gitlab.com/api/allyproject/commits') &lt;/execute_ipython&gt;</pre>
JSON Output	<pre>[ .....{   "id": "ed37a2f2",   "created_at": "2023-03-13T21:04:49.000-04:00",   "title": "Update README.md",   "message": "Update README.md",   "author": "SaptakS", }]</pre>

# Interacting

## ❑ Web Browsing

- ✔ Align with human habit
- ✔ Universal
- ✔ ...

- ✘ Limited action space
- ✔ ...

## ❑ Tool-based Interacting

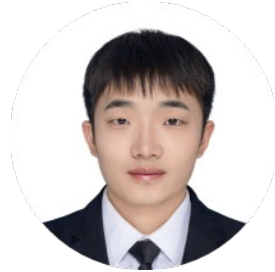
- ✔ More direct to interact with the webpage
- ✔ ...

- ✔ Some webpages may not support the tool-based interaction
- ✘ LFMs need to learn how to utilize the additional tool
- ✔ ...

# Towards Autonomous Web Interaction: Empowering WebAgents by LFM



Yujuan Ding



Liangbo Ning



Ziran Liang



Haohao Qu



Wenqi Fan



The Hong Kong Polytechnic University

June 9, 2026

