

Towards Next-Generation AI Agents for Web Automation with Large Foundation Models



Yujuan Ding¹



Liangbo Ning¹



Ziran Liang¹



Zhuohang Jiang¹



Haohao Qu¹



Wenqi Fan¹



Qing Li¹



Hui Liu²



Xiaoyong Wei¹



Philip S. Yu³

¹The Hong Kong Polytechnic University ²Michigan State University

³University of Illinois at Chicago



Website



Survey

August 4th (Day 2), 8:00 AM – 11:00 AM
Zoom ID: 816 7100 0487, Password: 123456

Tutorial Outline

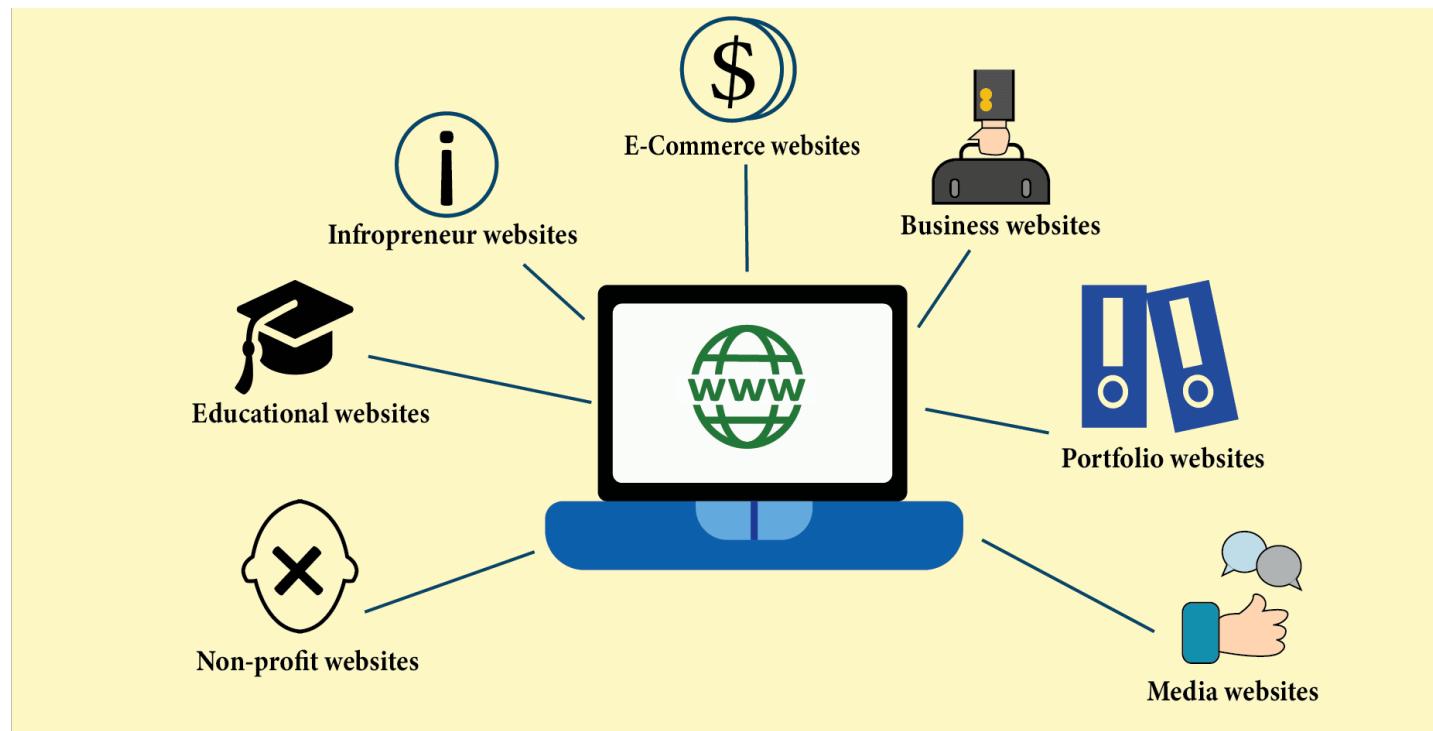
- **Part 1: Introduction of WebAgents (Yujuan Ding)**
- **Part 2: Preliminaries of AI Agents and LFM-based WebAgents (Zhuohang Jiang)**
- **Part 3: Architectures of WebAgents (Yujuan Ding)**
- **Coffee Break**
- **Part 4: Training of WebAgents (Yujuan Ding)**
- **Part 5: Trustworthy WebAgents (Haohao Qu)**
- **Part 6: Future directions of WebAgents (Zhuohang Jiang)**

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



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

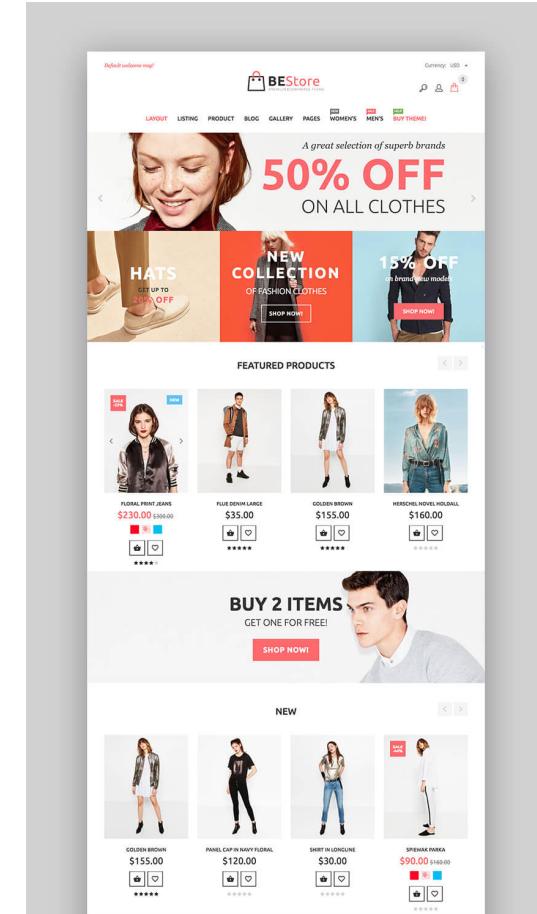
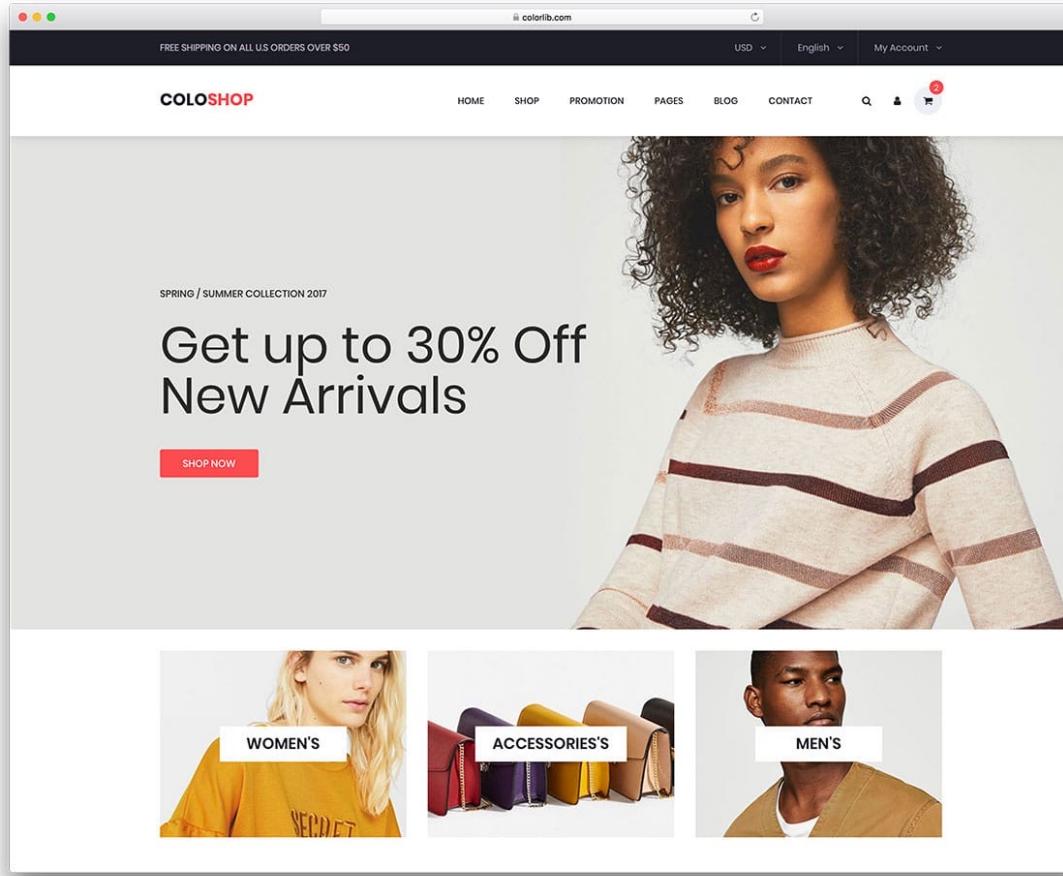
— Tim Berners-Lee,
Inventor of the World Wide Web,
ACM A.M. Turing Award Laureate 2016.



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.

Teaching & Learning

Generative AI e-learning Blog HKU Online Learning & MOOCs URFP Search...

> T&L@HKU > 4-Year UG Curriculum > Common Core > Academic Advising > Experiential Learning > Horizons > CICs > TALIC > Staff

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

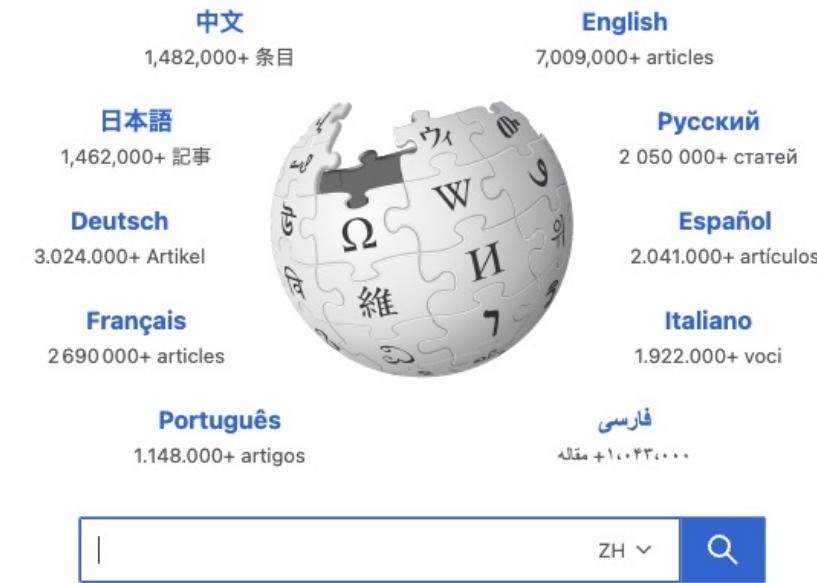
Professional Certificate Programs

FinTech (3 Courses)

FinTech Technologies I (3 Courses)

Advanced Cardiac Imaging (2 Courses)

WIKIPEDIA
自由的百科全书



Web for Healthcare



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



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#

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.



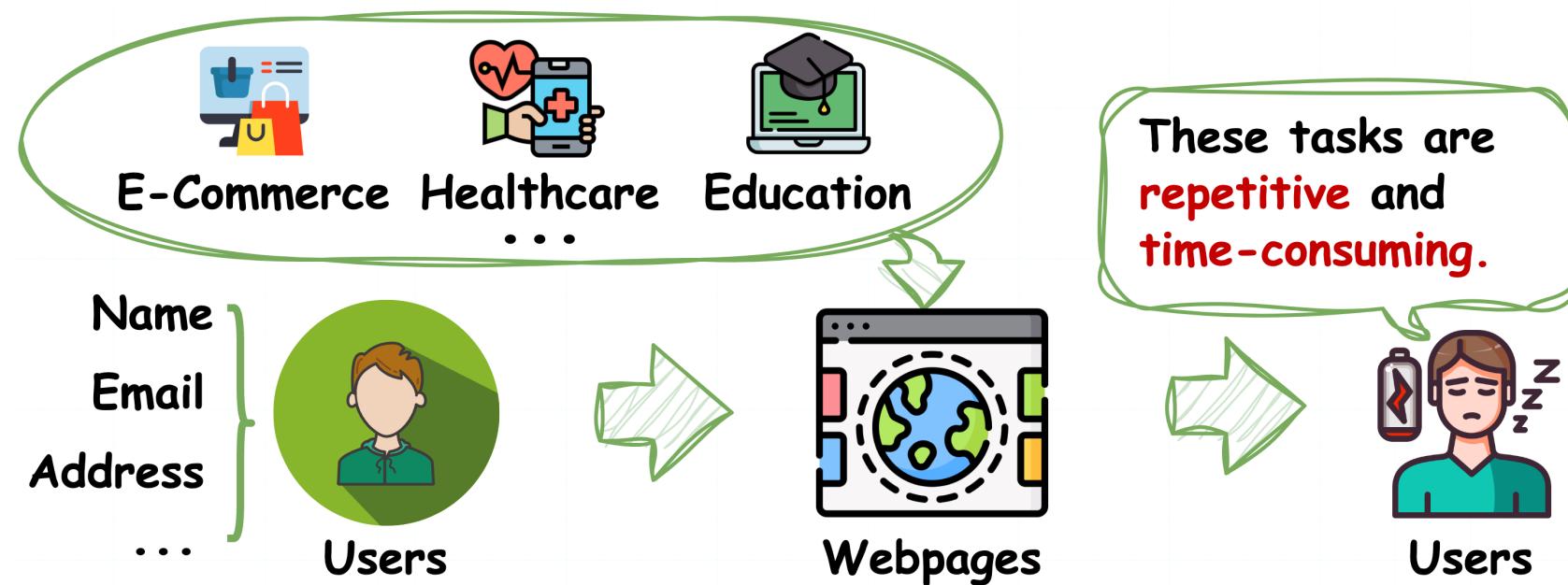
Online Doctor
Consult eDoctor Anytime Anywhere

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"

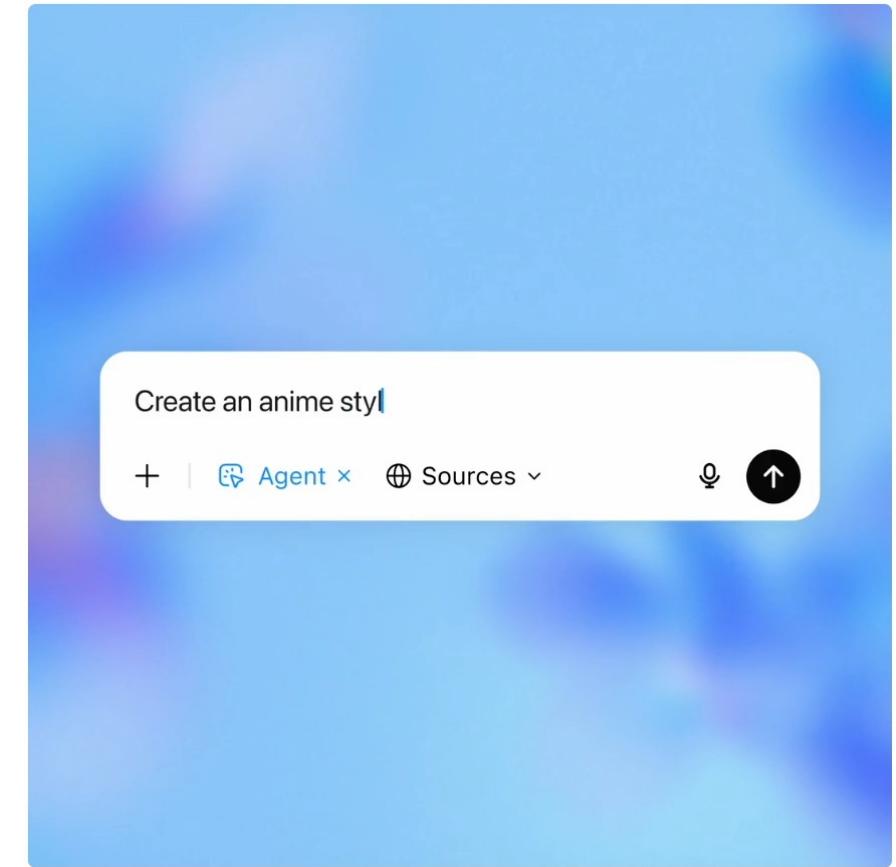
— Sam Altman

TED Ideas change everything



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

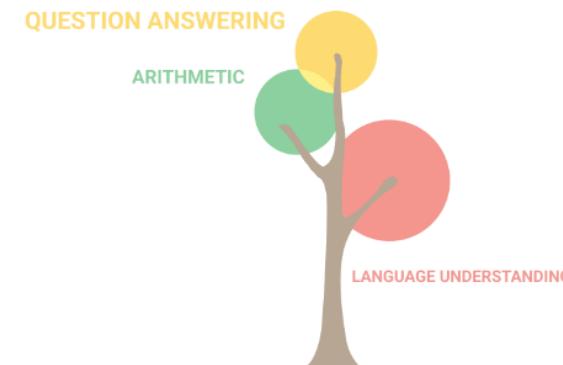
1,997,406 plays | Sam Altman | TED2025 • April 2025



Large Foundations Models (LFMs)

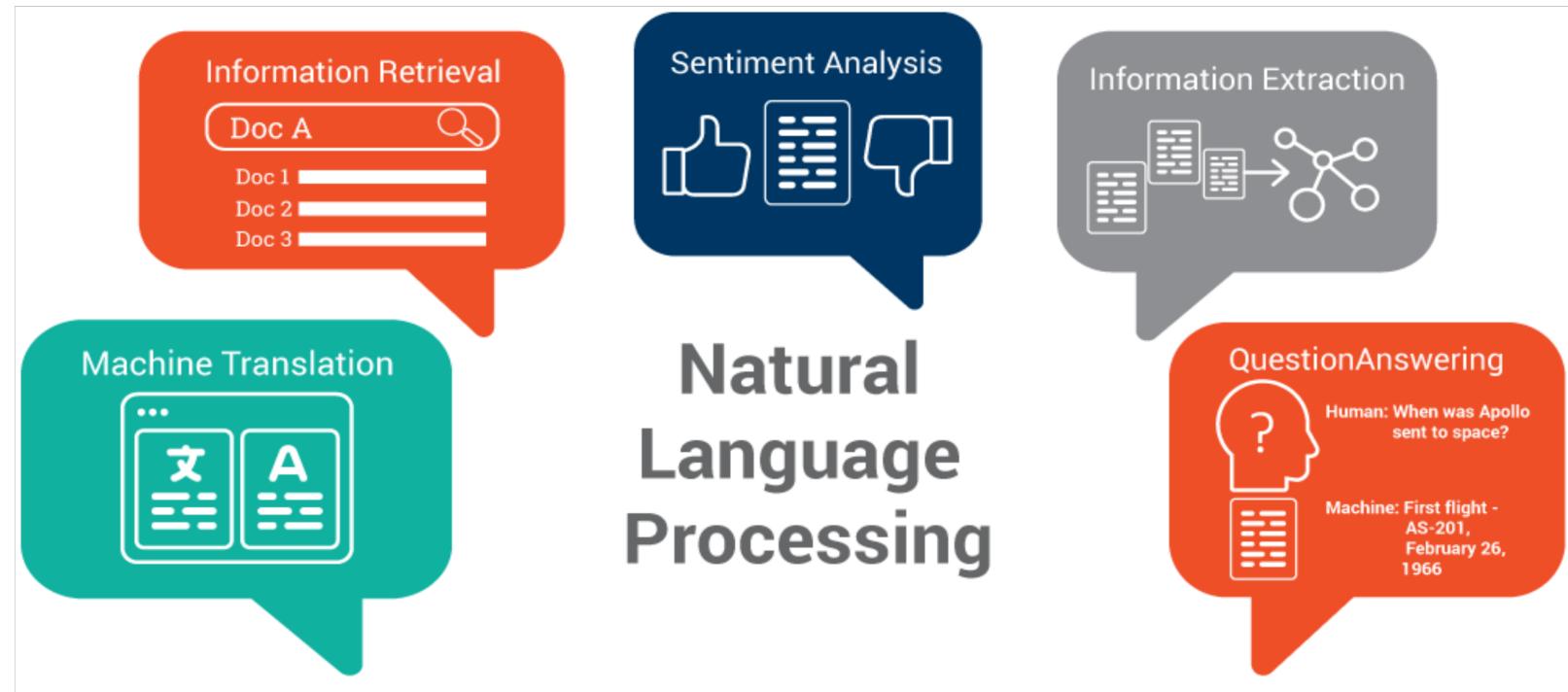


They Are Changing Our Lives !

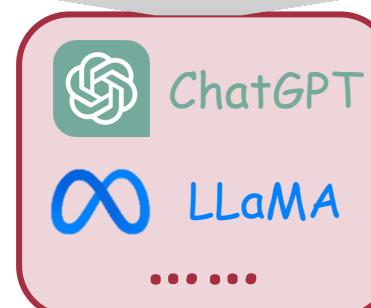


8 billion parameters

LFMs in Natural Language Processing



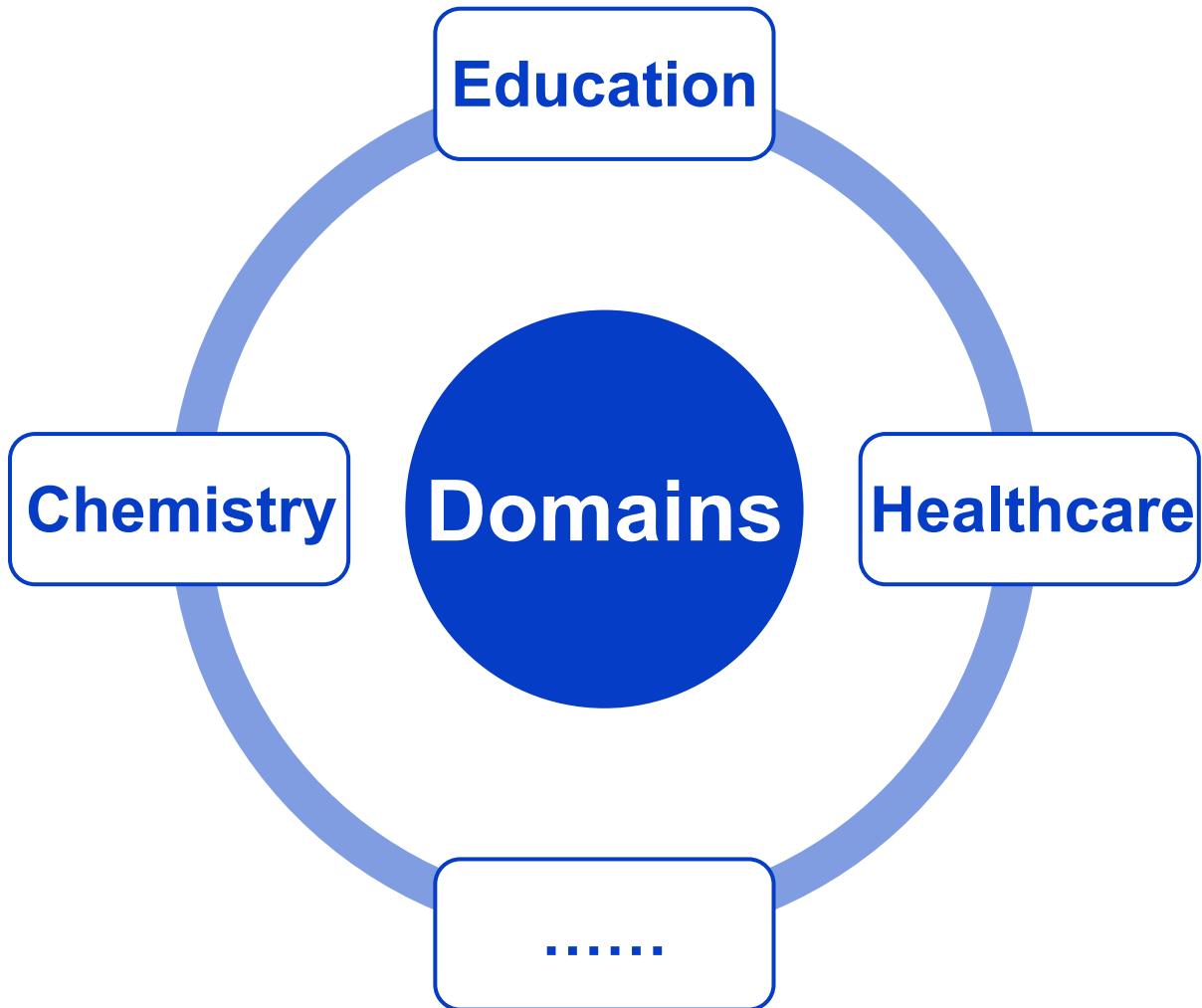
Input Text



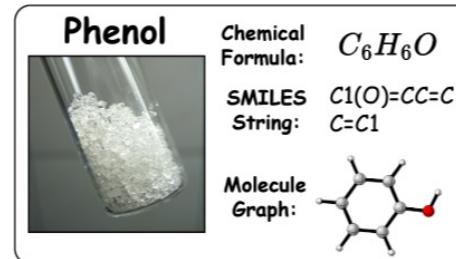
Generated Text

Large Language Models (LLMs)

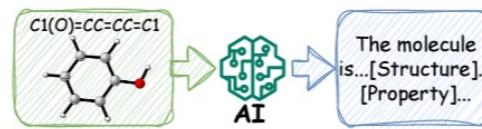
LFMs in Downstream Domains



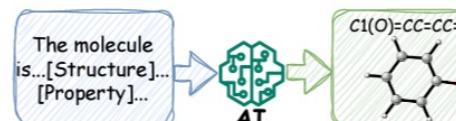
❑ Molecule discovery, etc.



(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.

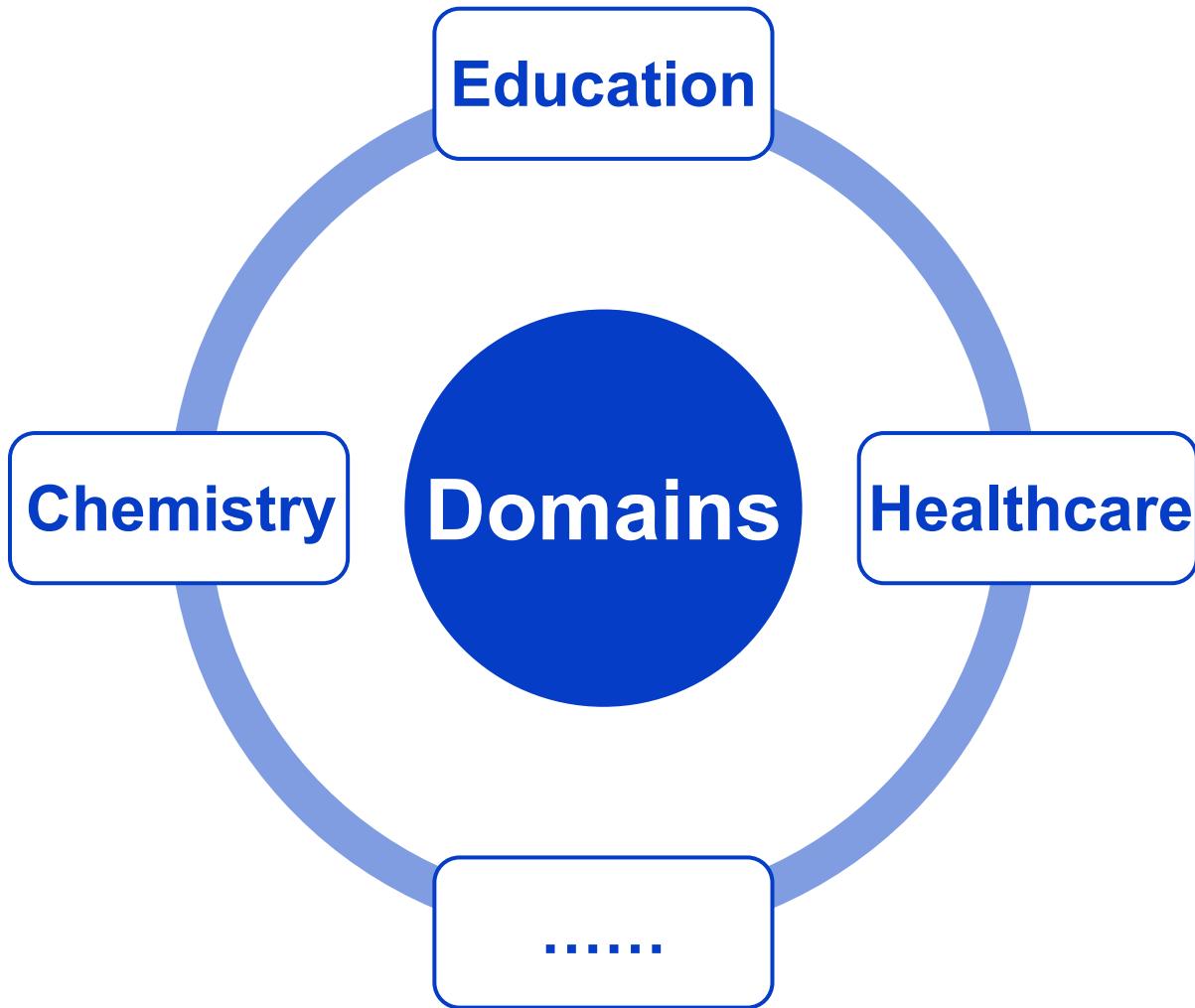
(b) Text-based Molecule Generation

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."

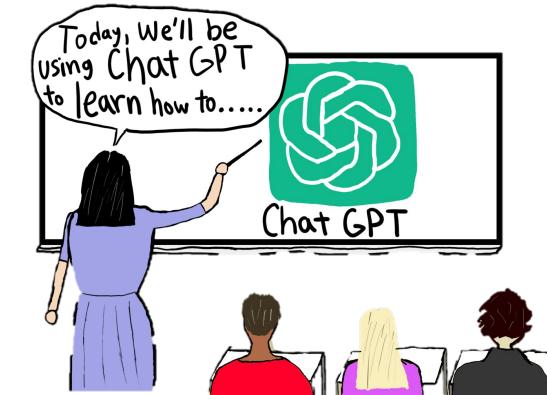
$C1=CC2=C(C(=C1)[O-])NC(=CC2=O)C(=O)O$

The interface shows a user inputting a molecule's SMILES string, receiving a detailed caption, and then generating a new molecule based on a descriptive text.

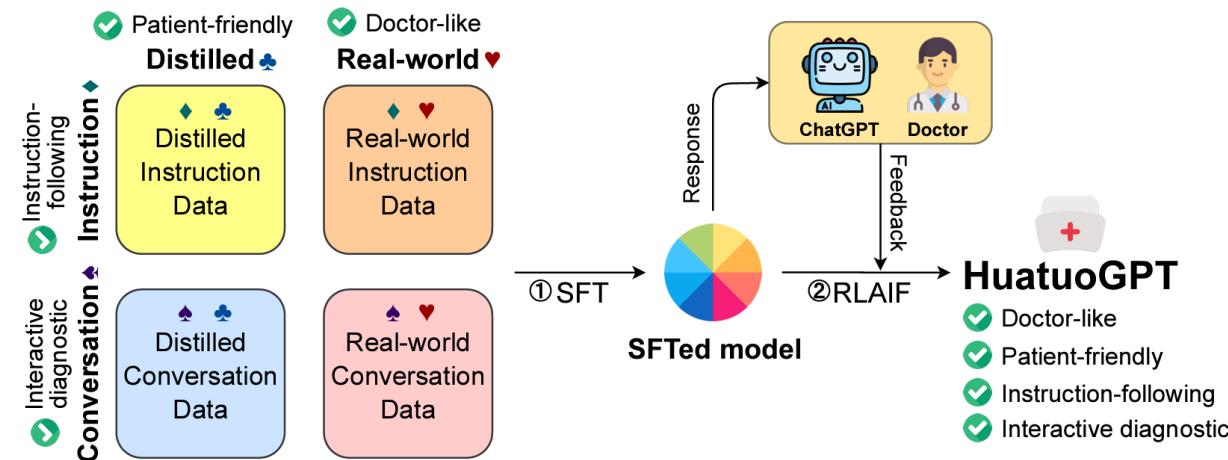
LFMs in Downstream Domains



Curriculum & Teaching, etc.



Medical consultation, etc.



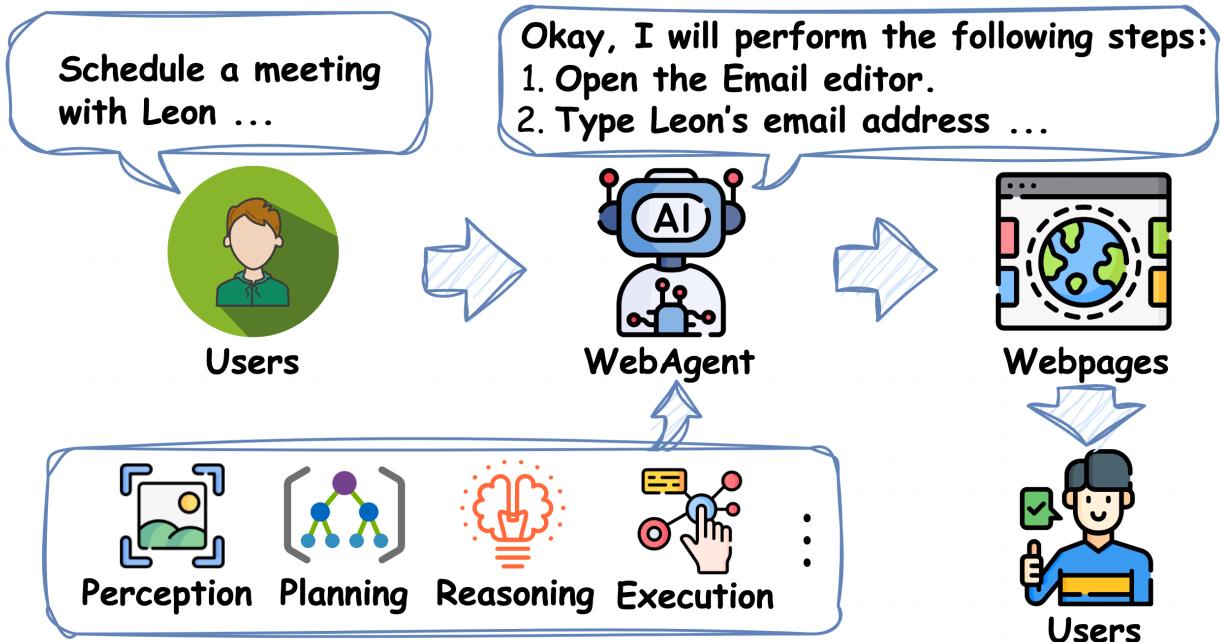
LFM-Empowered WebAgents



LFMs:

- **Understanding and Reasoning**
- **Action Generation and Execution**
- **Handling 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



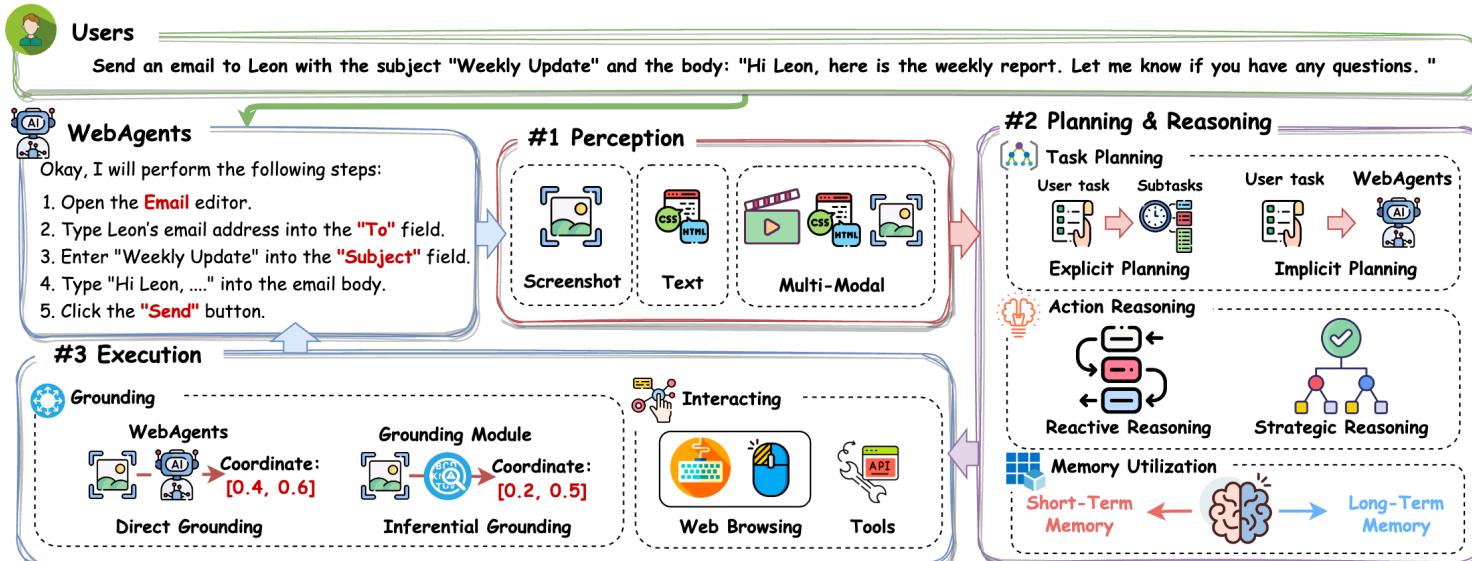
A Survey of WebAgents: Towards Next-Generation AI Agents for Web Automation with Large Foundation Models

Liangbo Ning¹, Ziran Liang¹, Zhuohang Jiang¹, Haohao Qu¹, Yujuan Ding¹, Wenqi Fan^{1*}, Xiao-yong Wei¹, Shanru Lin², Hui Liu³, Philip S. Yu⁴, Qing Li^{1*}

¹The Hong Kong Polytechnic University, ²City University of Hong Kong,

³Michigan State University, ⁴University of Illinois at Chicago

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



**Survey paper Tutorial
on KDD 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 (Yujuan Ding)
- Part 2: Preliminaries of AI Agents and LFM-based WebAgents (Zhuohang Jiang)
- Part 3: Architectures of WebAgents (Yujuan Ding)
- Coffee Break
- Part 4: Training of WebAgents (Yujuan Ding)
- Part 5: Trustworthy WebAgents (Haohao Qu)
- Part 6: Future directions of WebAgents (Zhuohang Jiang)

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



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



Presenter
Zhuohang Jiang
HK PolyU

- 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

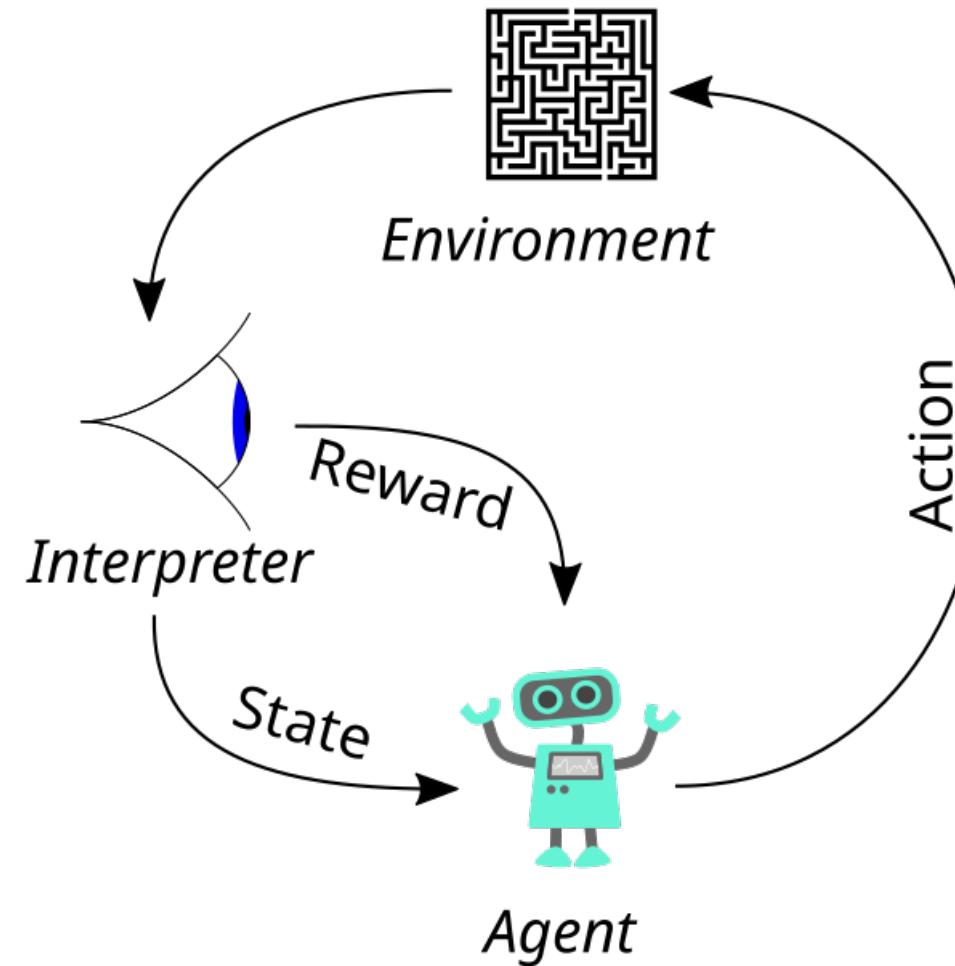


Website of this tutorial

- **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

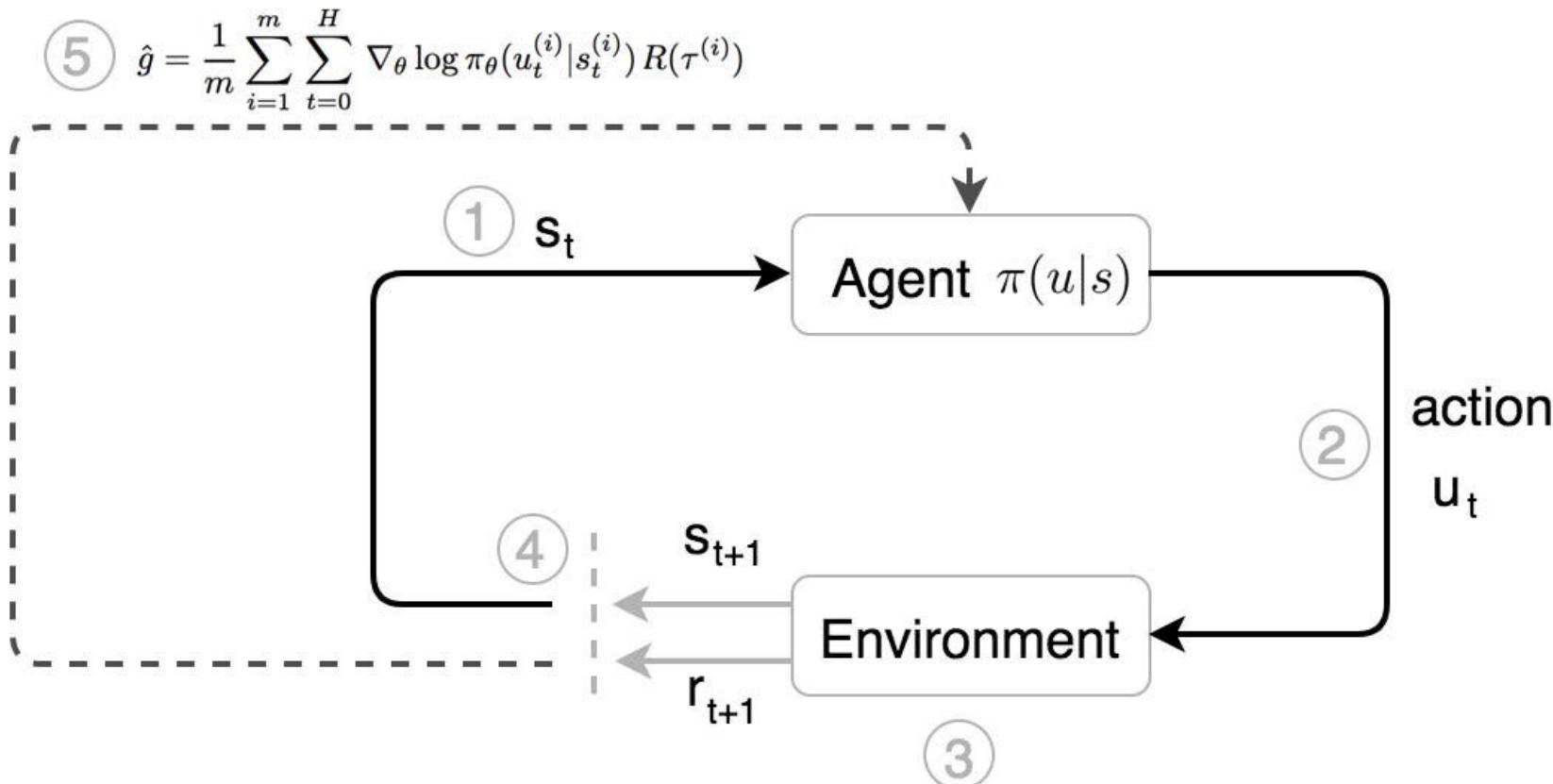
- Q-Learning learns the optimal action-value function by iteratively updating Q-values based on rewards and future estimates.



Reinforcement Learning

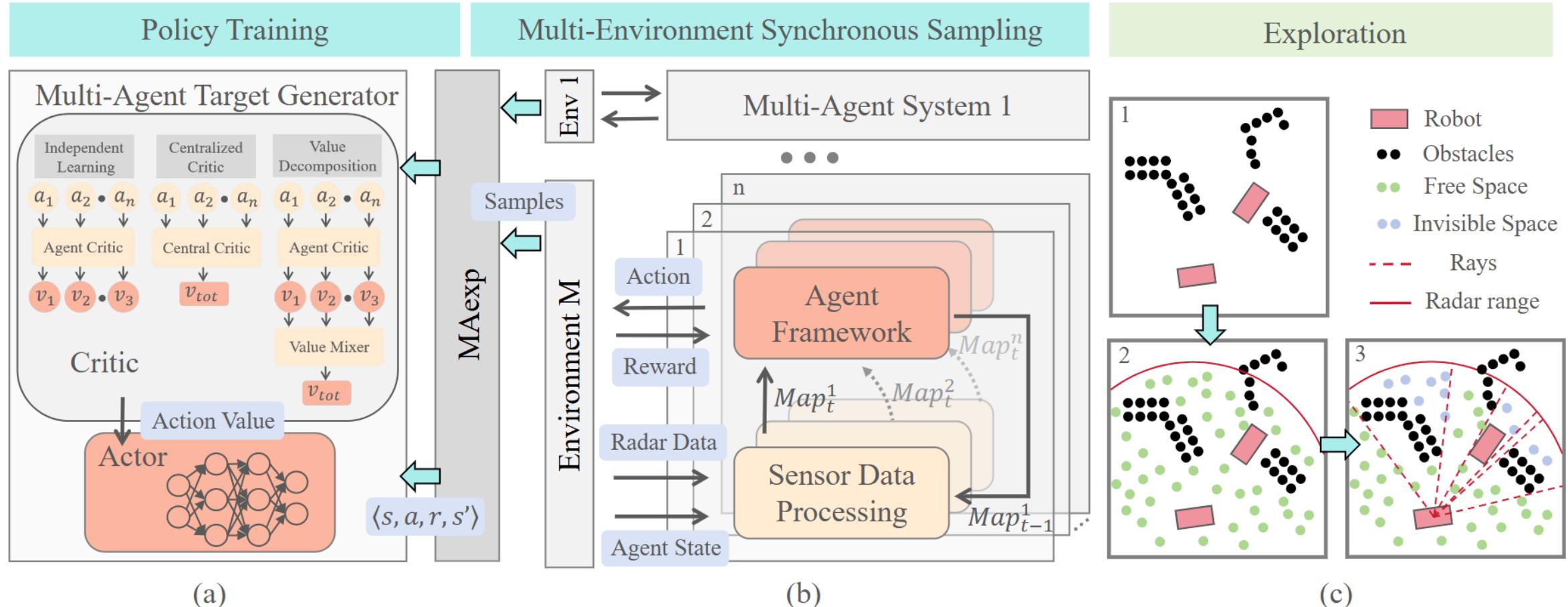


- Policy Gradient directly optimizes actions by maximizing expected rewards.



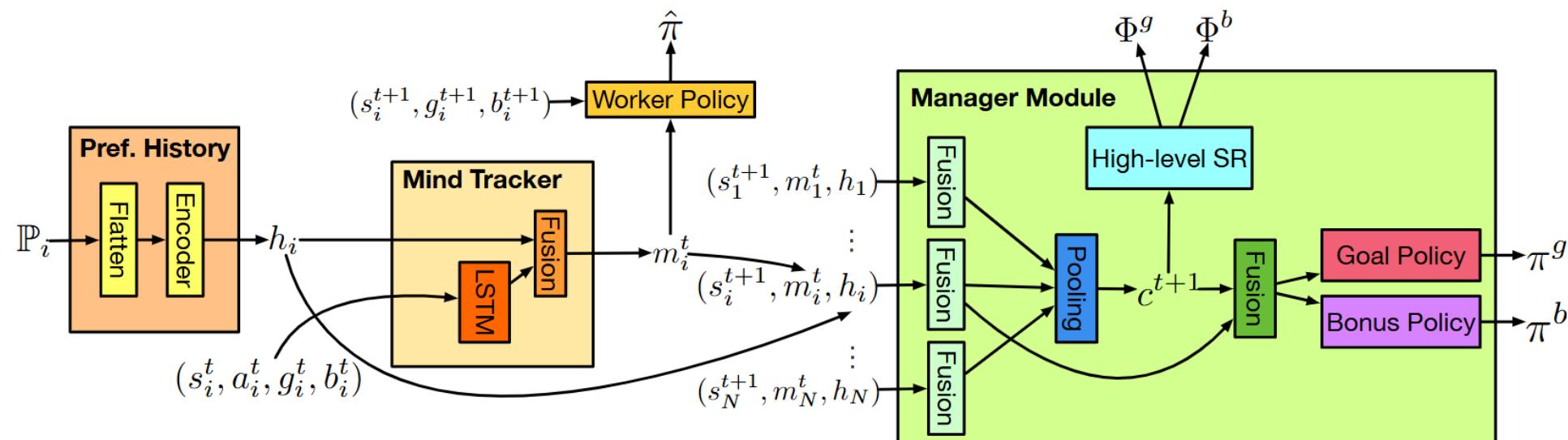
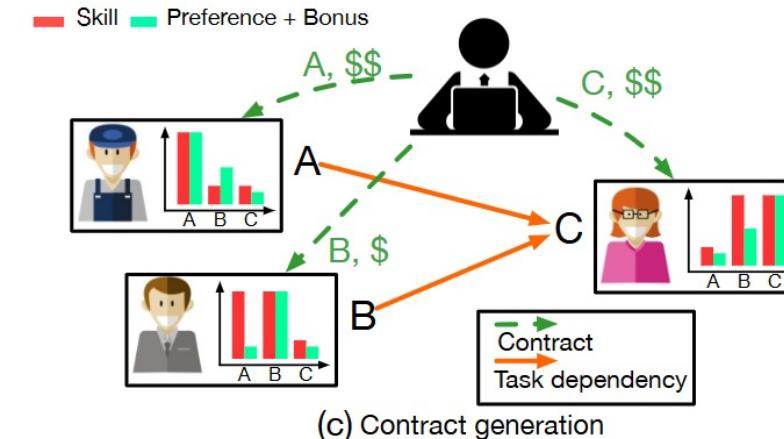
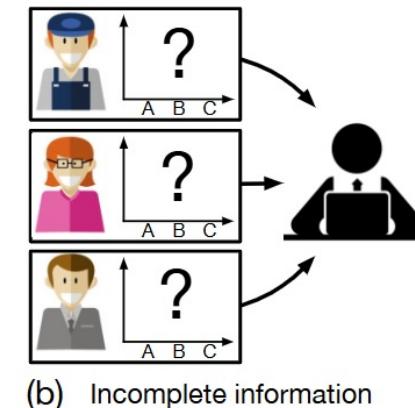
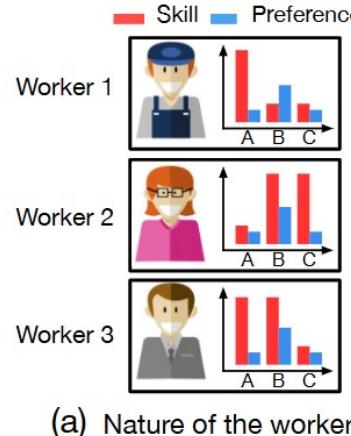
RL-based Agents

□ Exploration (Maexp)



RL-based Agents

□ Management (M³RL)

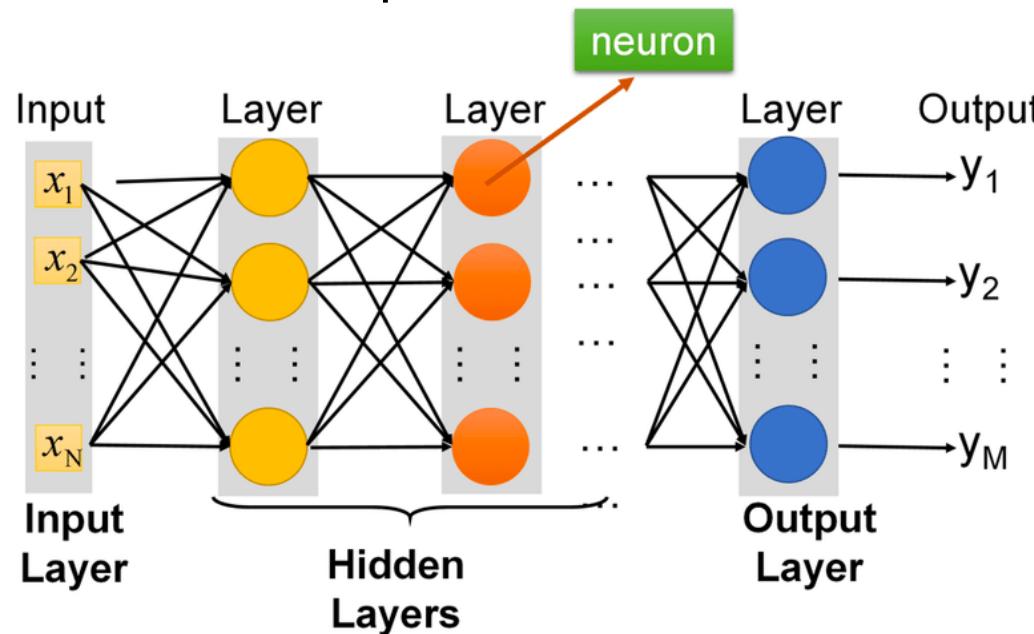


Deep Reinforcement Learning

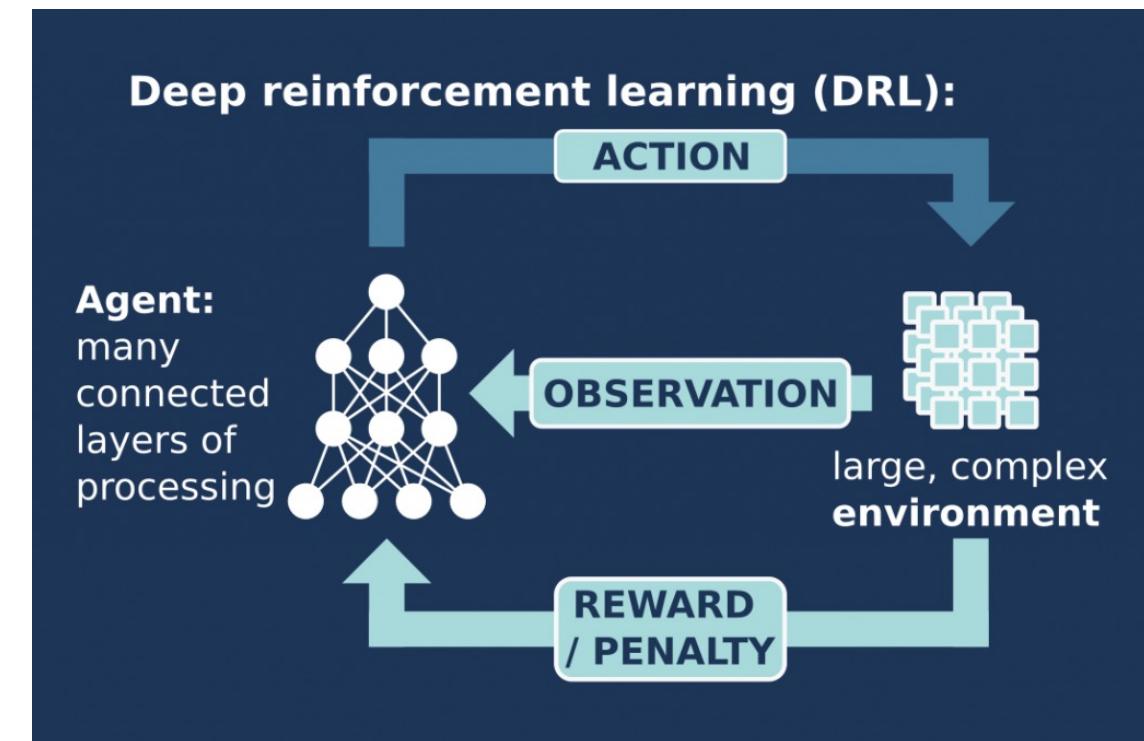


- The success of deep learning has led to **Deep Reinforcement Learning (DRL)**, combining neural networks with reinforcement learning for complex decision-making.

Deep Neural Network



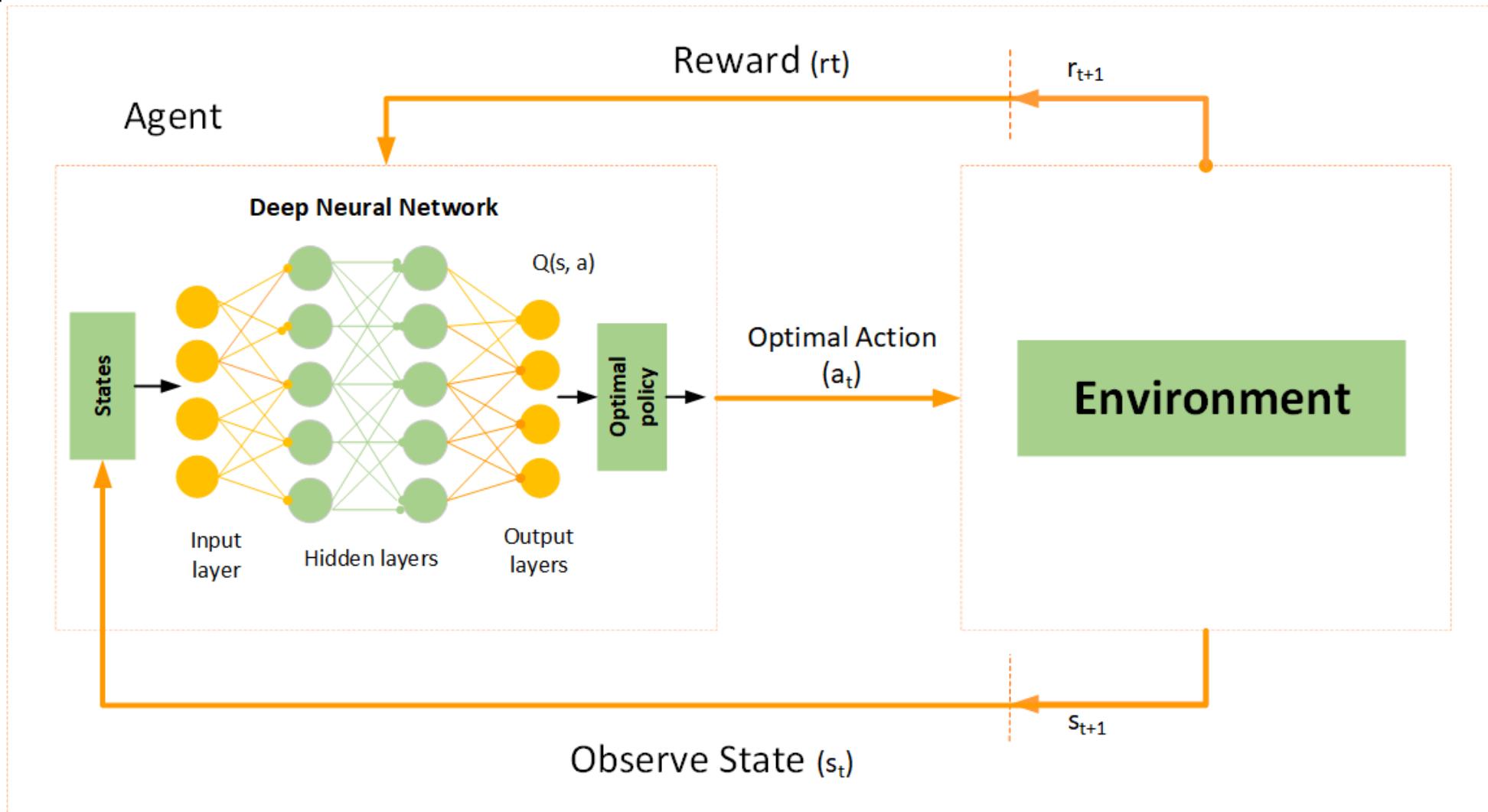
Deep Reinforcement Learning



Deep Reinforcement Learning



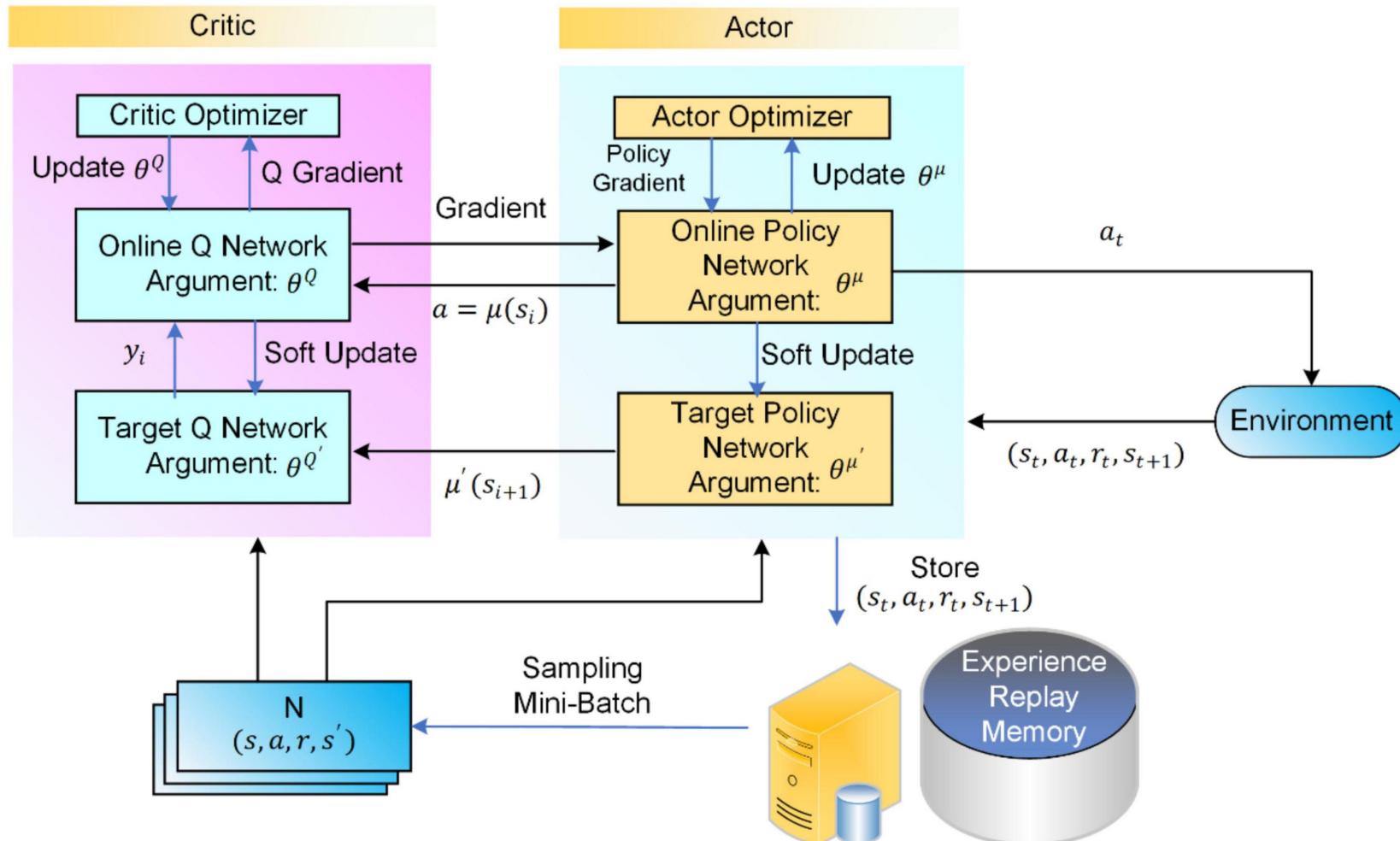
□ Deep Q Network



Deep Reinforcement Learning



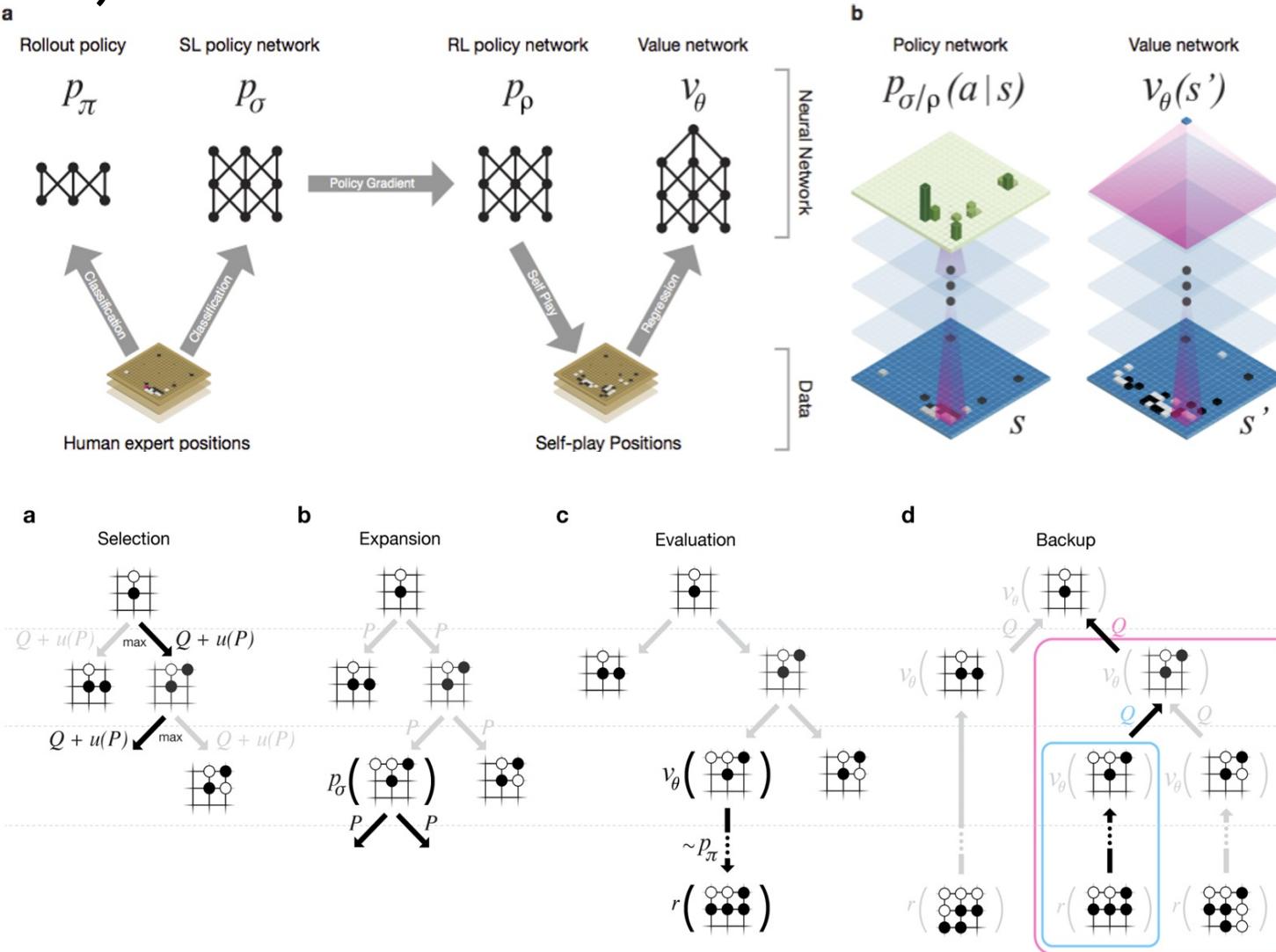
□ Deep Deterministic Policy Gradient



Deep Reinforcement Learning Agents



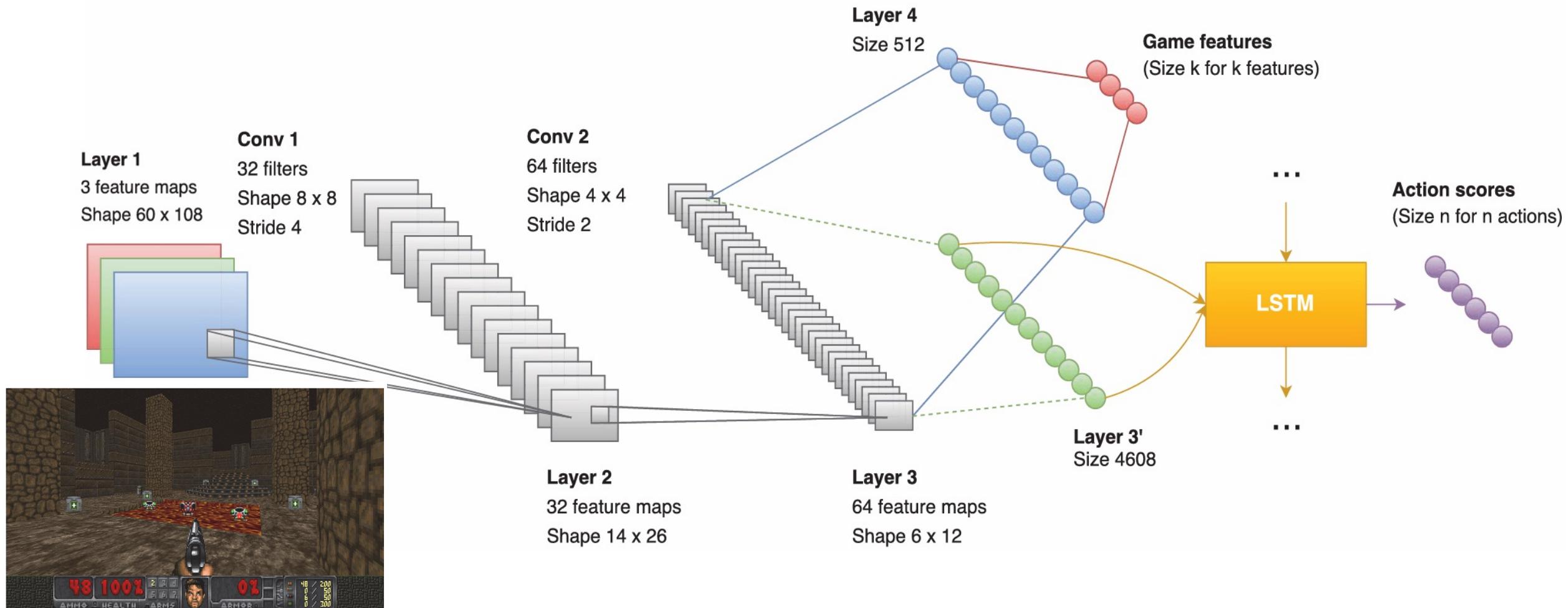
□ Game (AlphaGo)



Deep Reinforcement Learning Agents



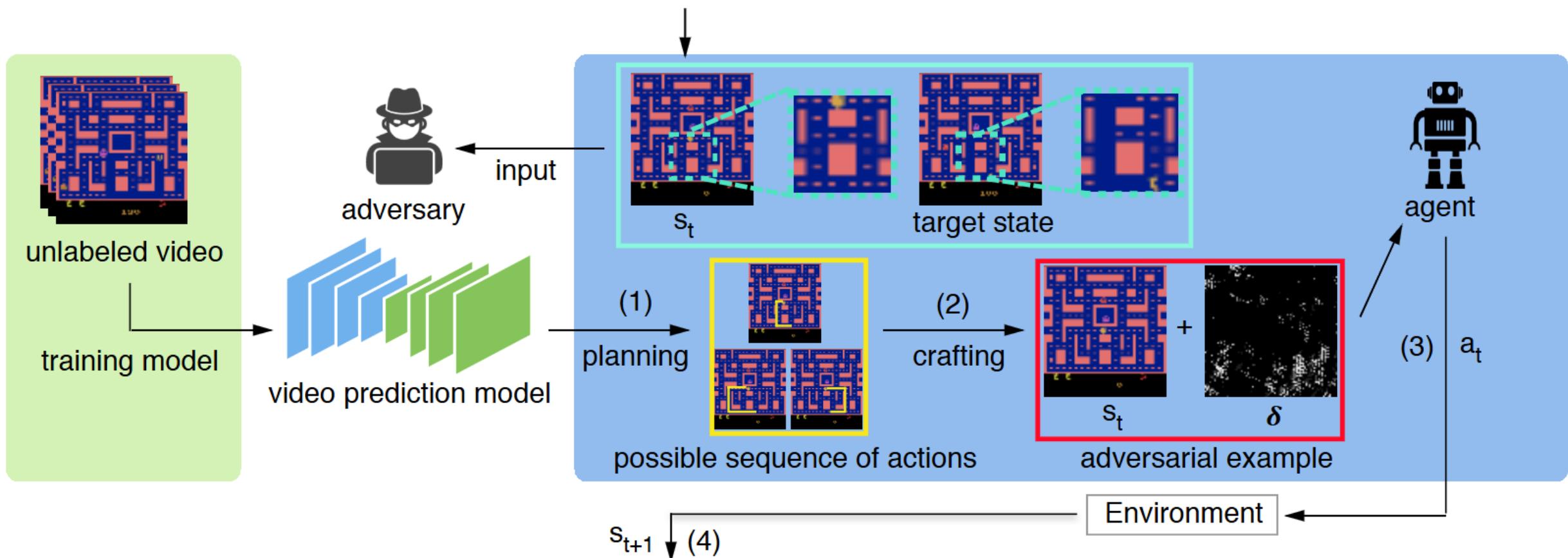
□ Game (Playing FPS Games with Deep Reinforcement Learning)



Deep Reinforcement Learning Agents



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



Applications and Challenges



☐ Applications

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



Go games

☐ Challenges

- ✖ Lack of World Knowledge
- Lack of Adaptability
- ...



Finance

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

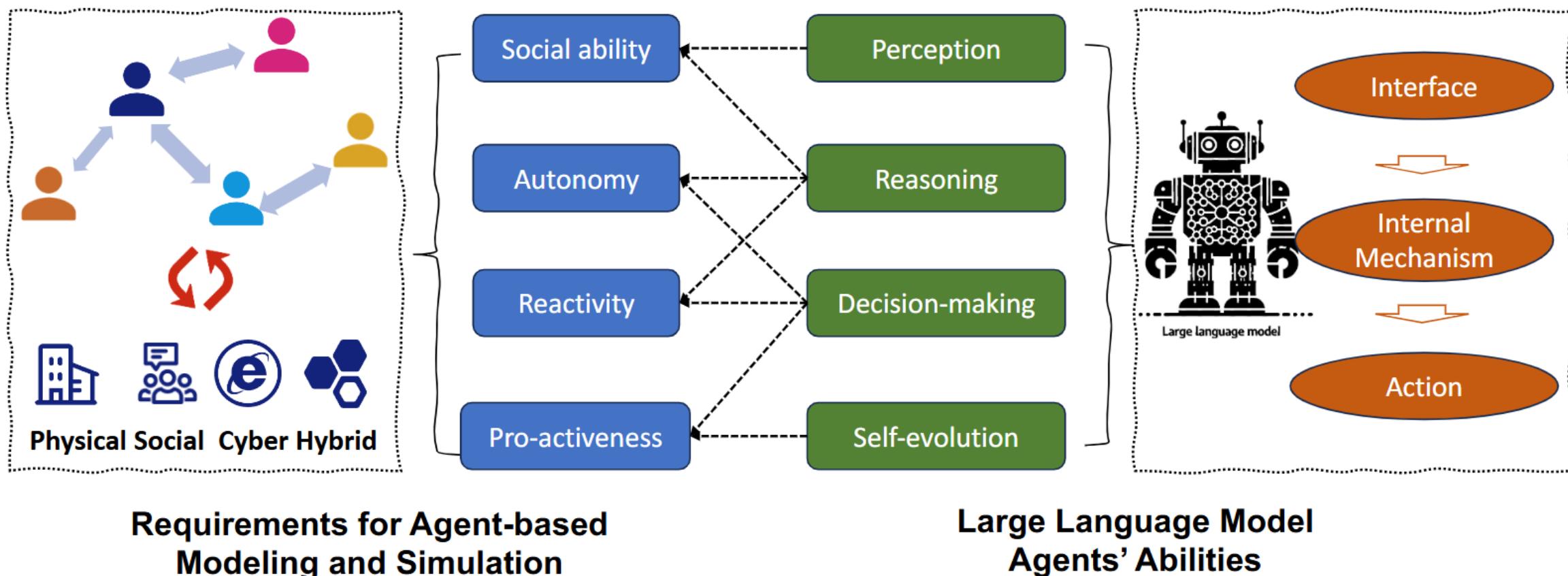


Website of this tutorial

- 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

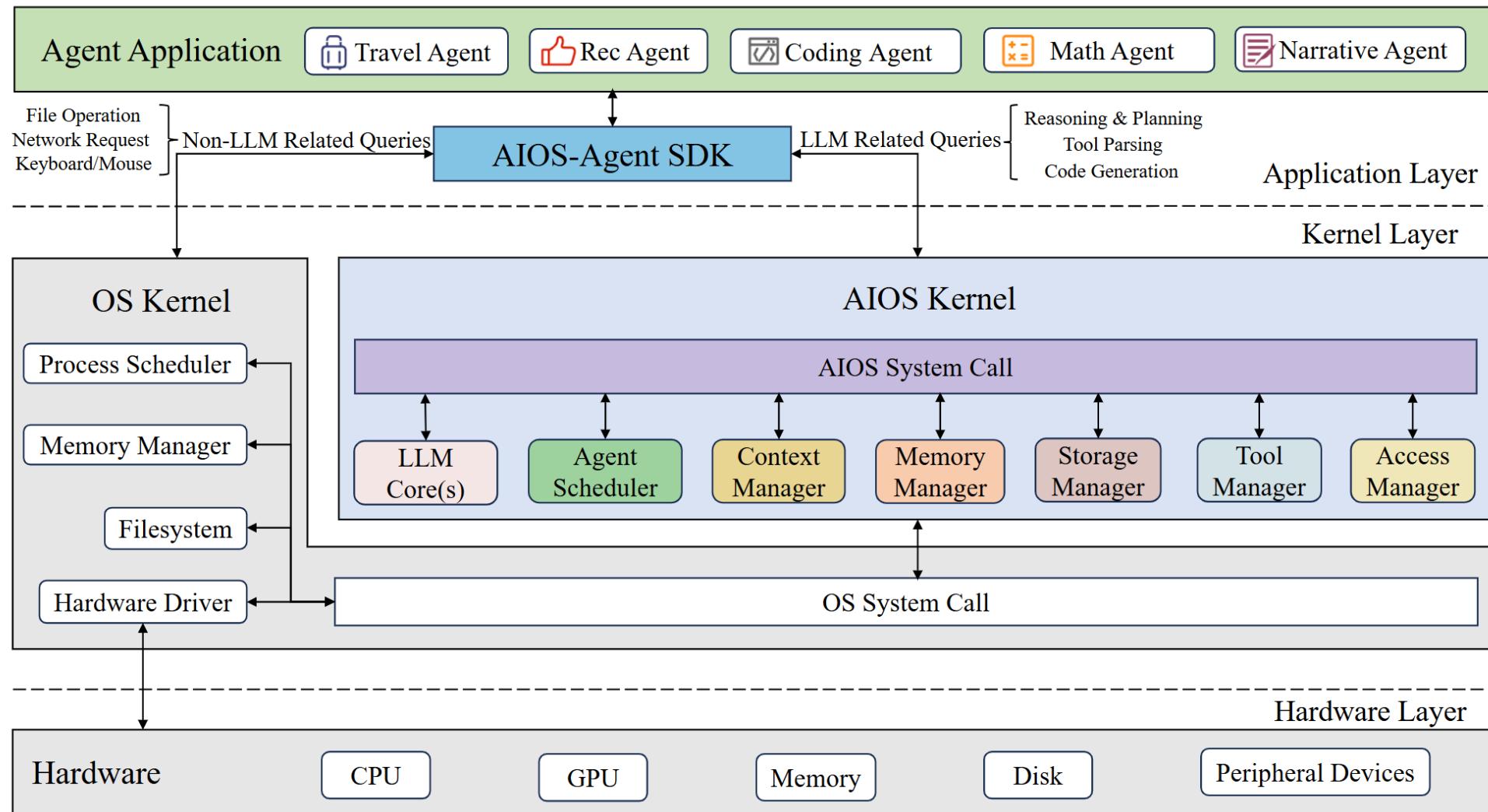
- Large Foundation Models with billion-level parameters have demonstrated remarkable intelligence characterized by rich intrinsic knowledge.



Applications and Challenges



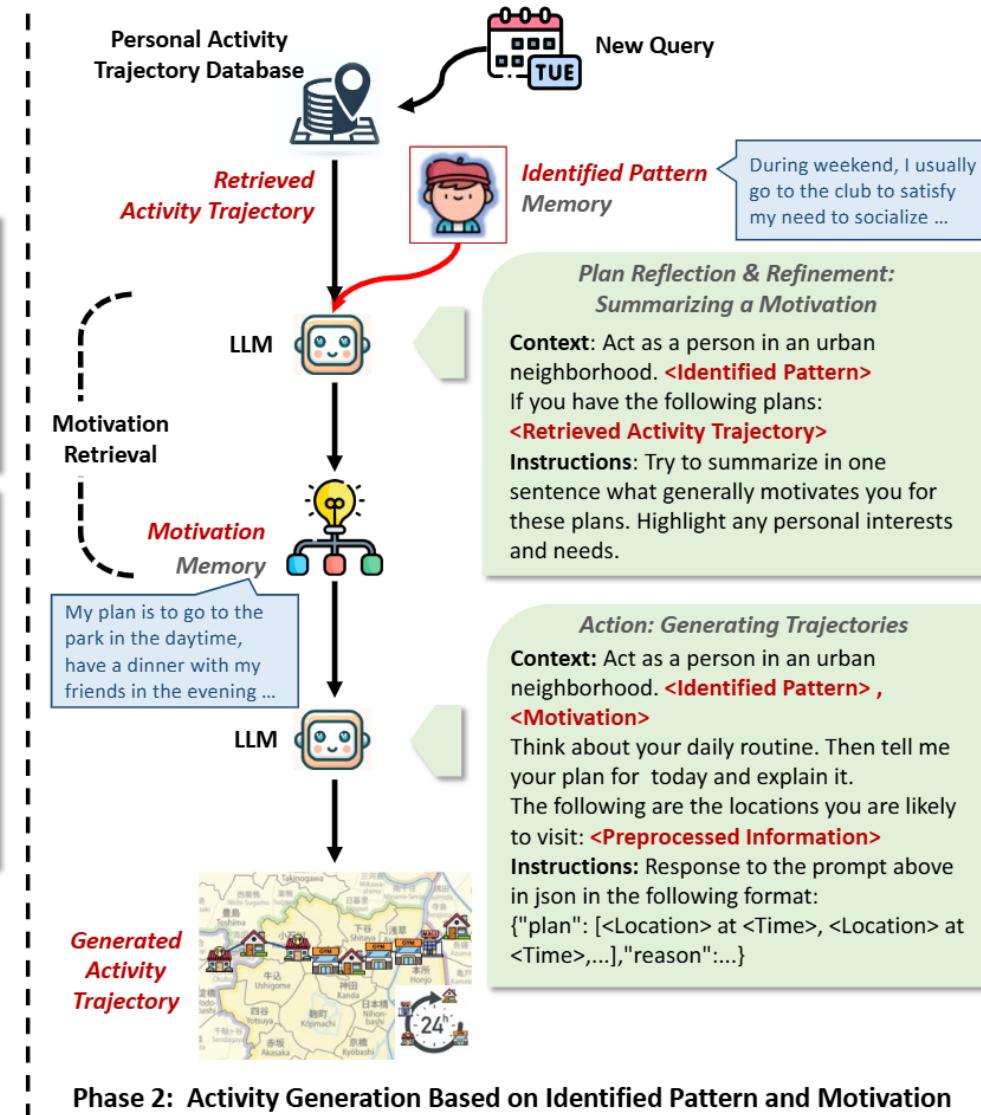
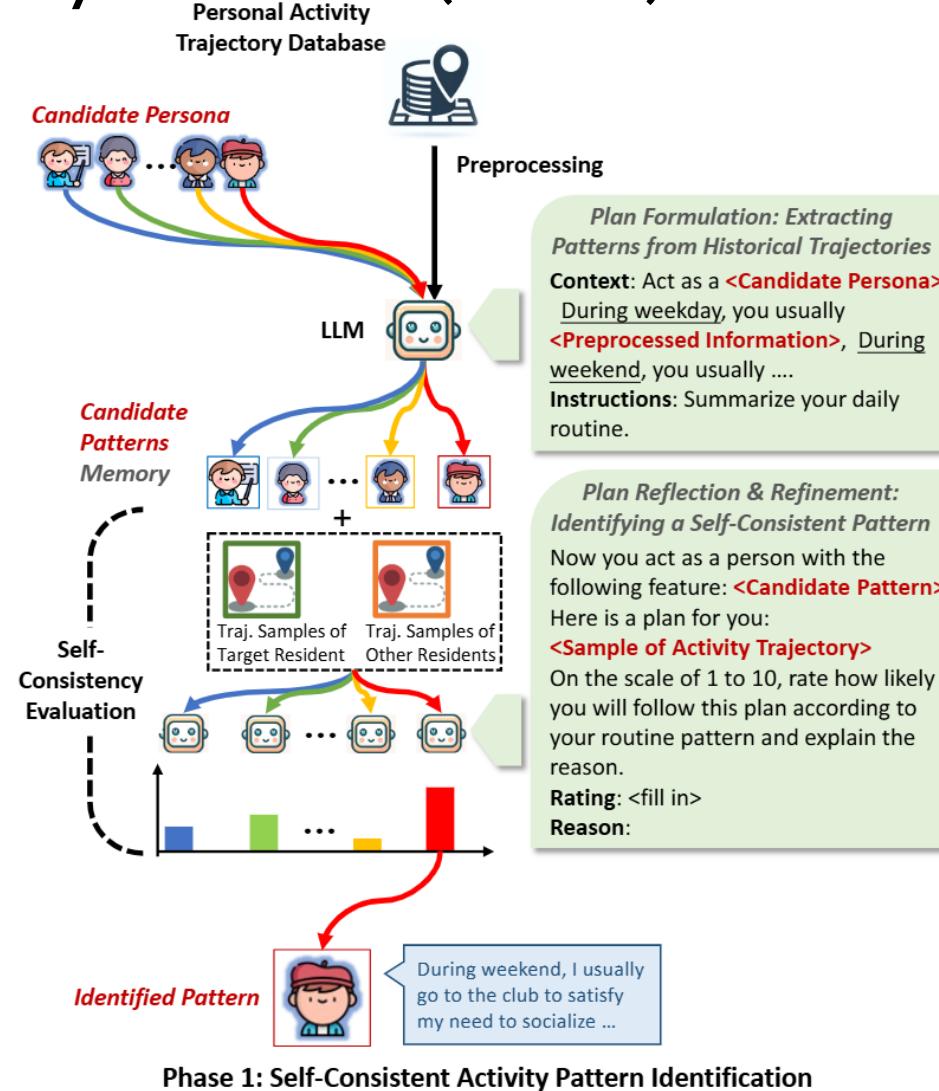
□ OS (AIOS)



Applications and Challenges



❑ Mobility Generation (LLMob)



Applications and Challenges

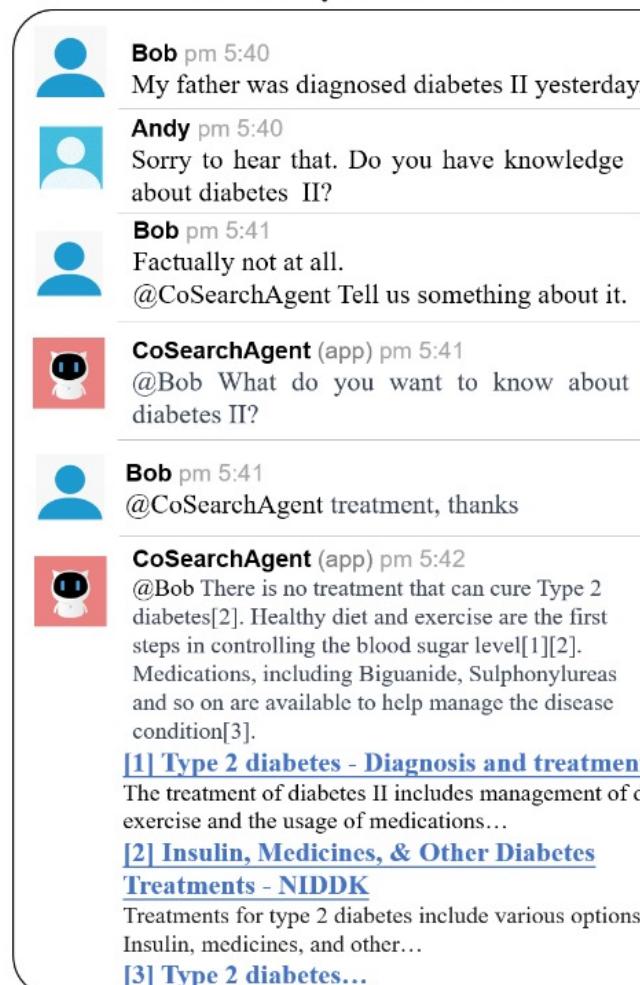
□ World Model (WorldCoder)



Applications and Challenges

□ Web Search (CoSearchAgent)

Multi-Party Conversation



Stage I: Query Processing

Rewrite Query Rewrite the query based on the conversational context.

Output Tell us something about **diabetes II**.

Ask Clarify Question Ask question about the ambiguous parts in the query.

Output What do you want to know about diabetes II?

Rewrite Query Rewrite the query based on the conversational context.

Output Tell us **treatment** about **diabetes II**.

Ask Clarify Question Ask question about the ambiguous parts in the query.

Output No clarify question.

Stage III: Retrieval-Augmented Generation

Retrieval-Augmented Generation Respond to user queries based on references

Output There is no treatment that can cure Type 2 diabetes[2]. Healthy diet and exercise are the first steps in controlling the blood sugar level[1][2]. Medications, including Biguanide, Sulphonylureas and so on are available to help manage the disease condition[3].

Stage II: Result Presentation

Fetch Retrieval search results relevant to user queries and extract the text contents from them.

Query Tell us treatment about diabetes II.

[1] Type 2 diabetes - Diagnosis and treatment

<h2>Treatment</h2>
<p>Management of type 2 diabetes includes:</p>
Healthy eating.
...
<p>These steps make it more likely that ...</p>

[1] Type 2 diabetes - Diagnosis and treatment

Treatment
Management of type 2 diabetes includes:
Healthy eating.
...
These steps make it more likely that blood ...

Extract Extract the portions of text content relevant to the query and summarize them into a reference collection.

[1] Type 2 diabetes - Diagnosis and treatment

Treatment
The treatment of diabetes II includes management of diet, exercise and the usage of medications.

[1] Type 2 diabetes - Diagnosis and treatment

Treatment
Management of type 2 diabetes includes:
Healthy eating.
...
These steps make it more likely that blood ...

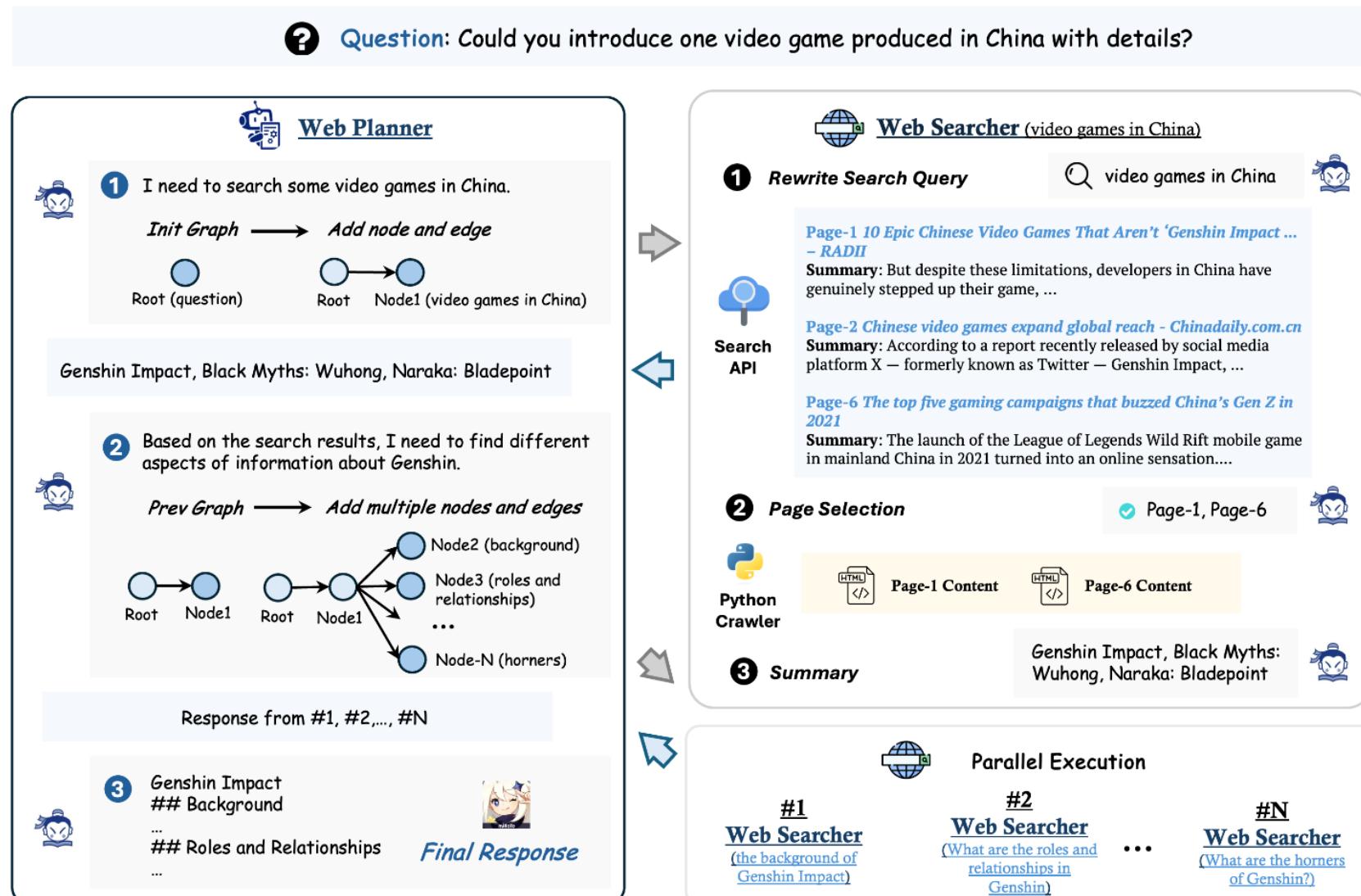
Users

CoSearchAgent Framework

Applications and Challenges

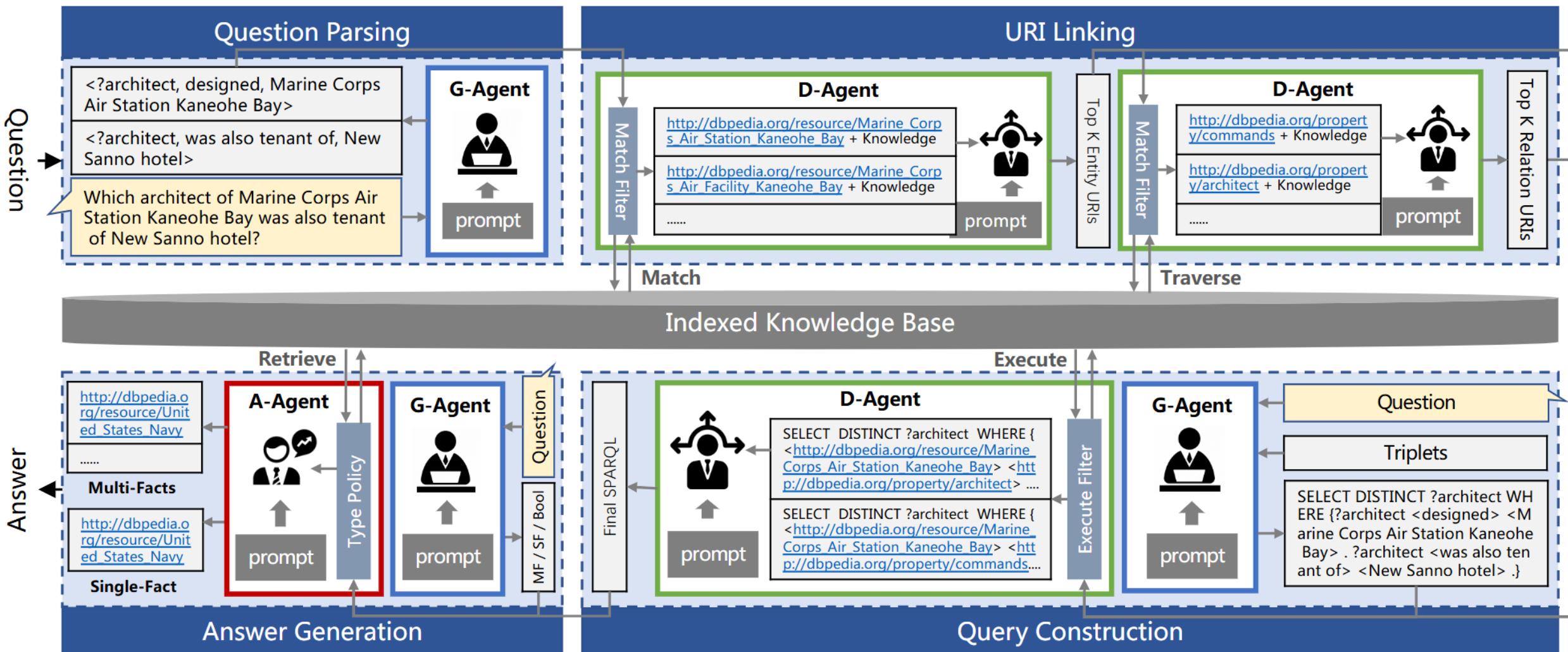


□ Web Search (MindSearch)



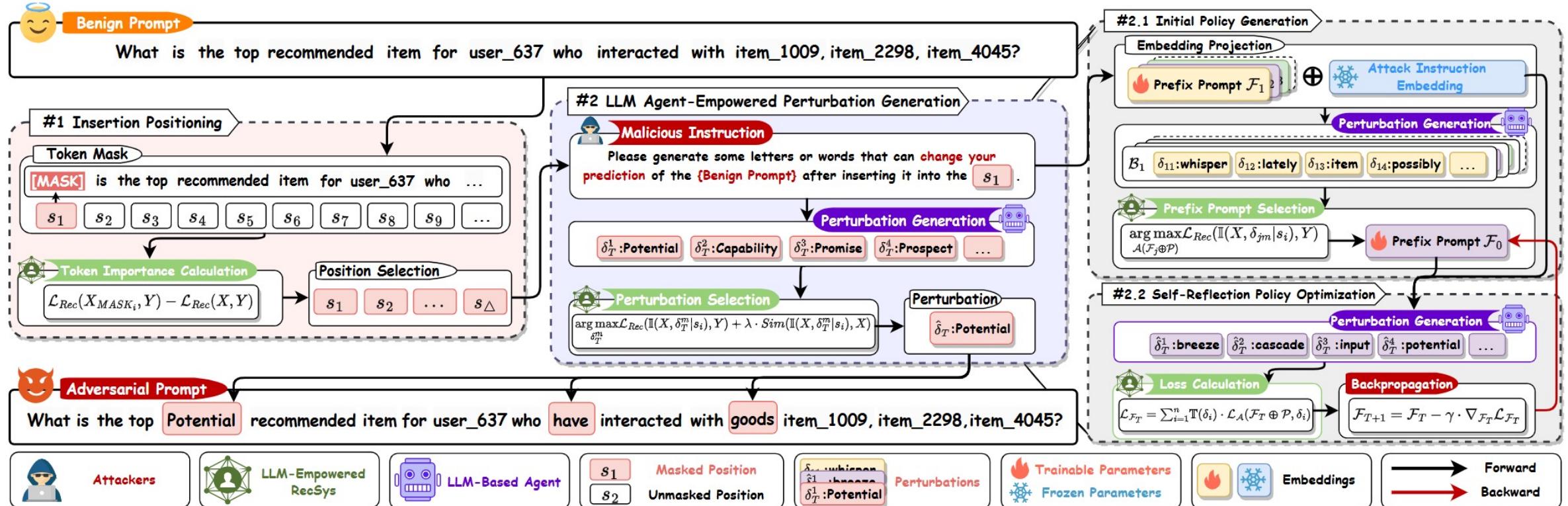
Applications and Challenges

□ KBQA (Triad)



Applications and Challenges

□ RecSys Attack (CheatAgent)



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



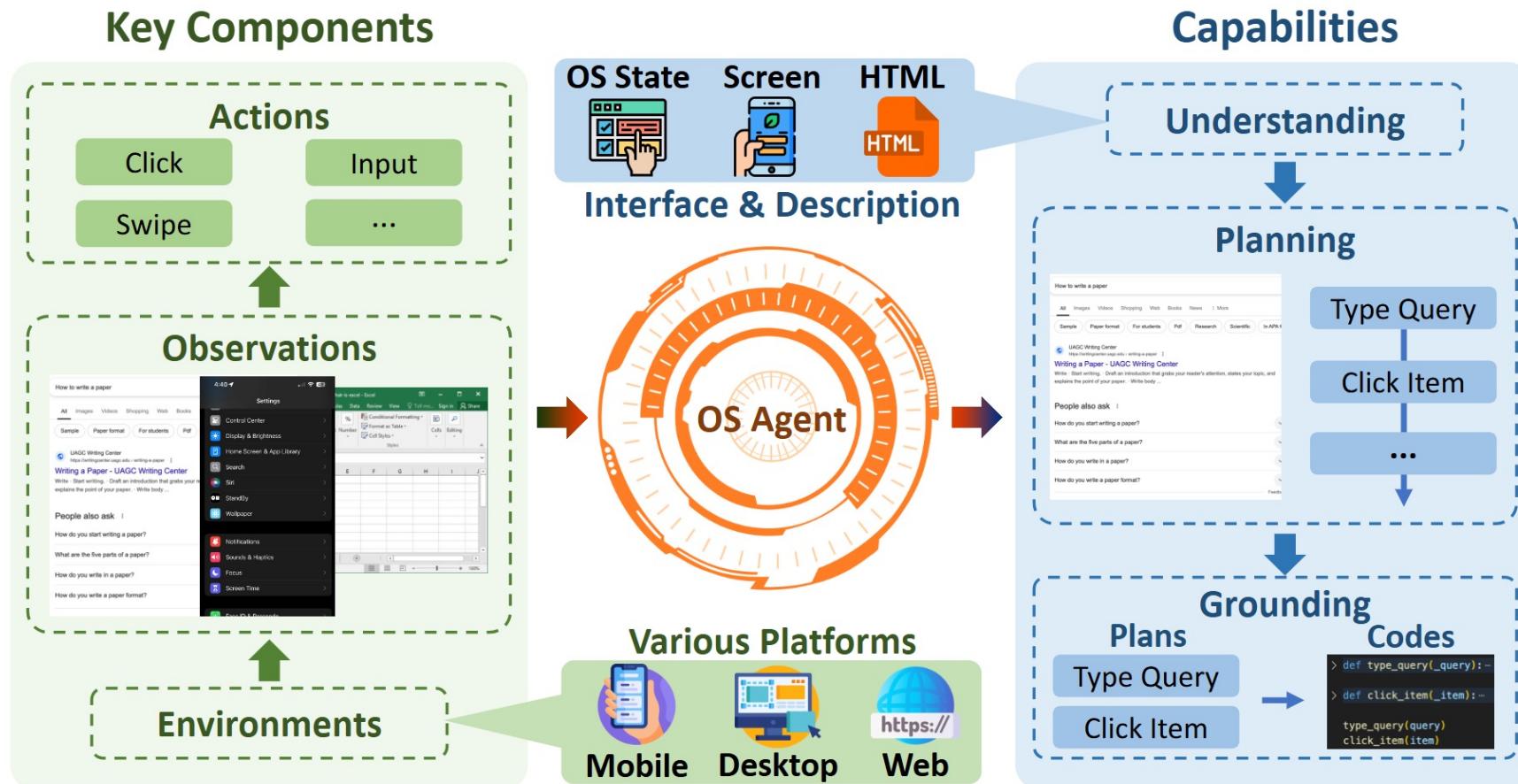
Website of this tutorial

- 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

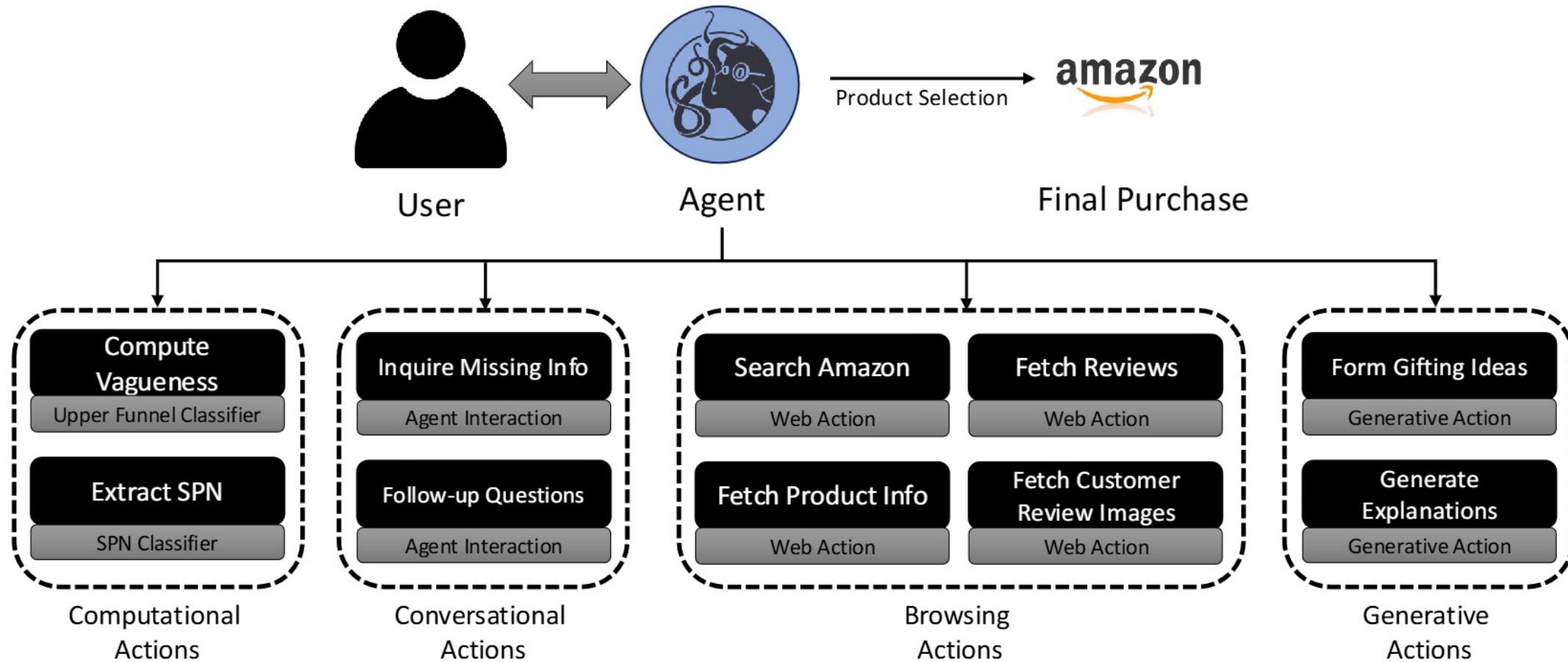


- WebAgents aim to perceive the environment and reason over user instructions to generate and execute actions step-by-step toward fulfilling the user's goal.



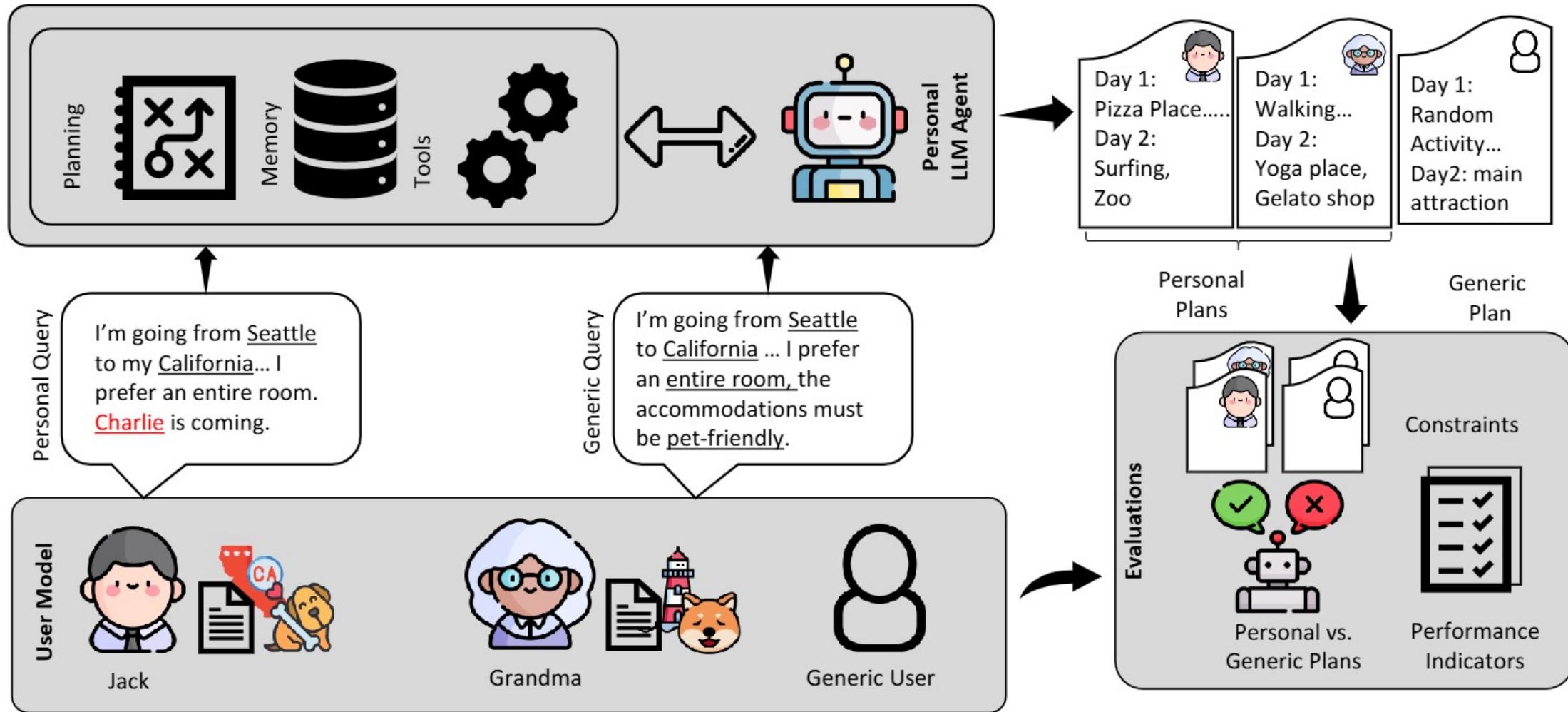
Applications and Challenges

□ Shopping (SPN Shopping Agent)



Applications and Challenges

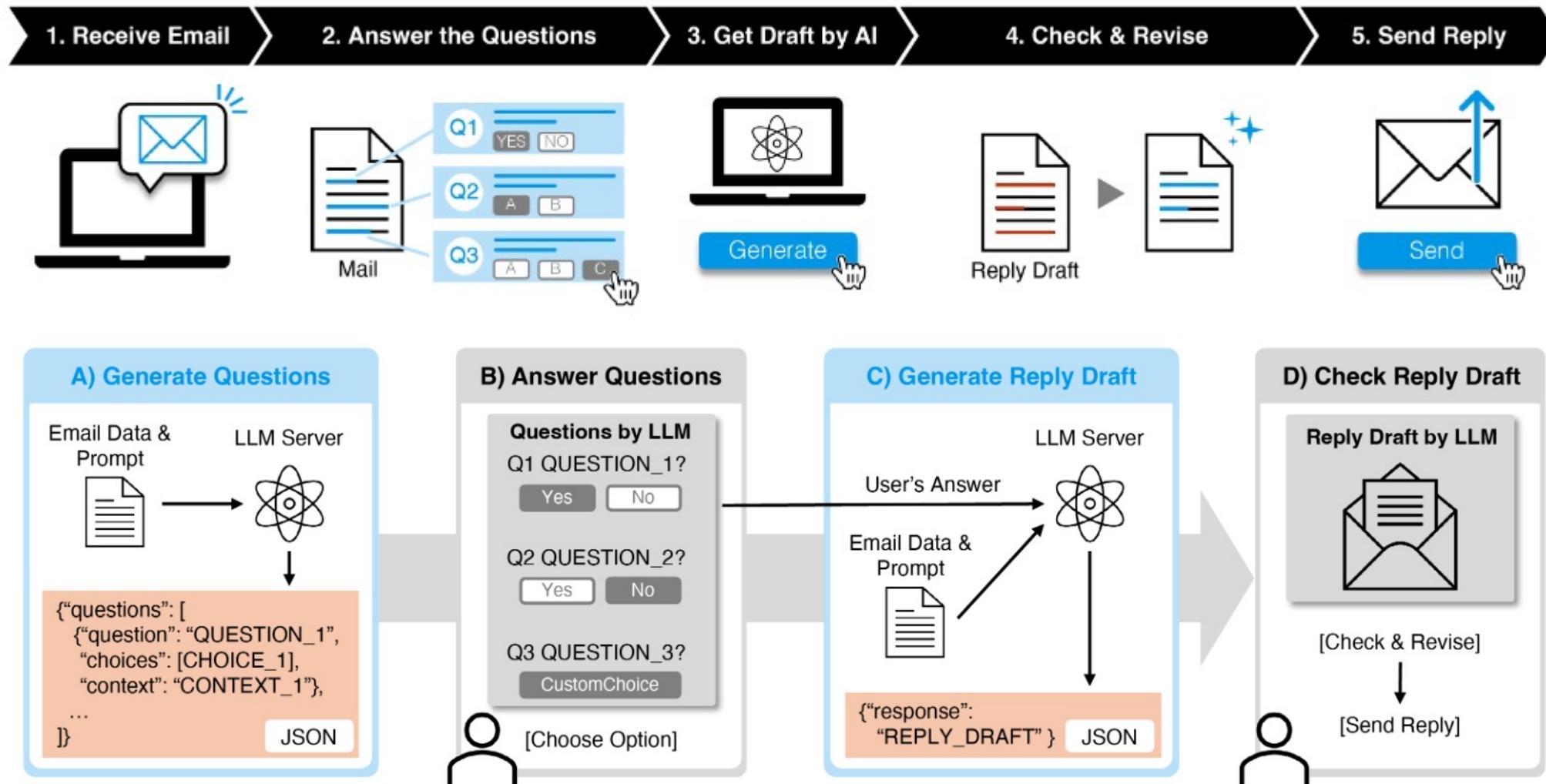
□ Travel Planning (TravelPlanner+)



Applications and Challenges



□ Email Reply(ResQ)



Tutorial Outline

- Part 1: Introduction of WebAgents (Yujuan Ding)
- Part 2: Preliminaries of AI Agents and LFM-based WebAgents (Zhuohang Jiang)
- Part 3: Architectures of WebAgents (Yujuan Ding)
- Coffee Break
- Part 4: Training of WebAgents (Yujuan Ding)
- Part 5: Trustworthy WebAgents (Haohao Qu)
- Part 6: Future directions of WebAgents (Zhuohang Jiang)

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



PART 3: Architectures of WebAgents



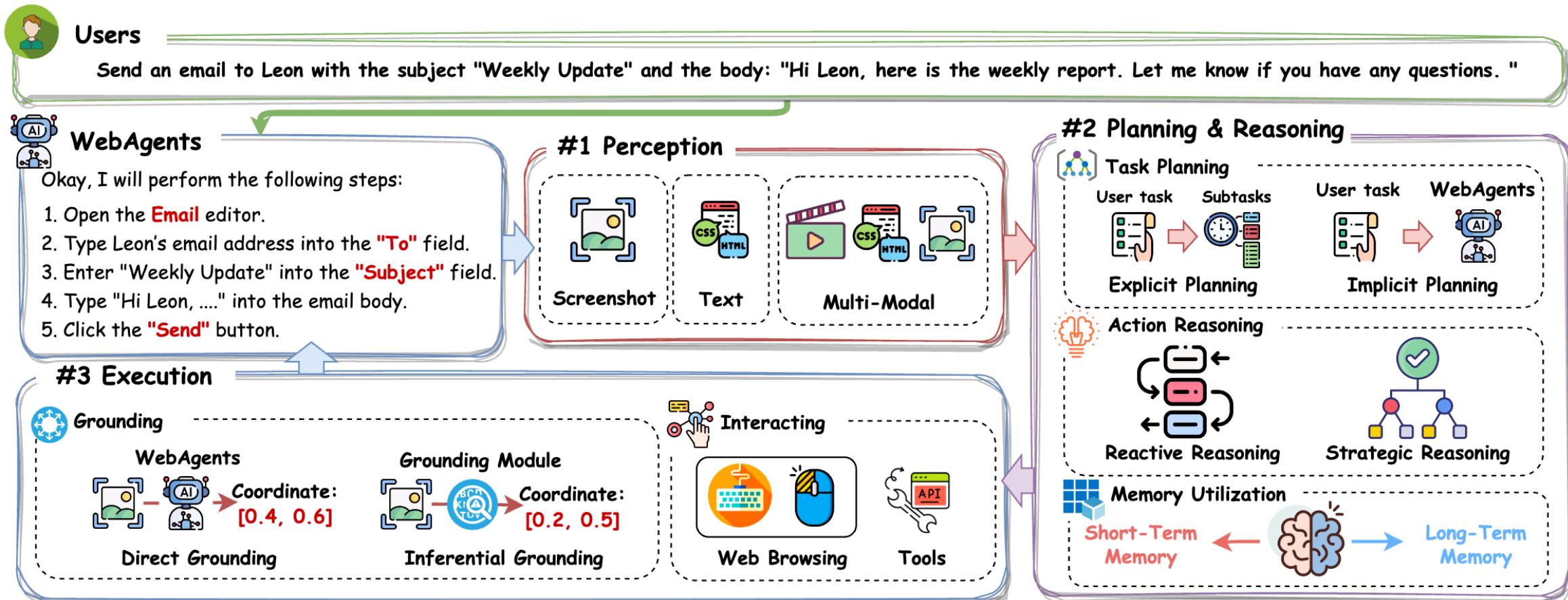
Presenter
Yujuan Ding
HK PolyU

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

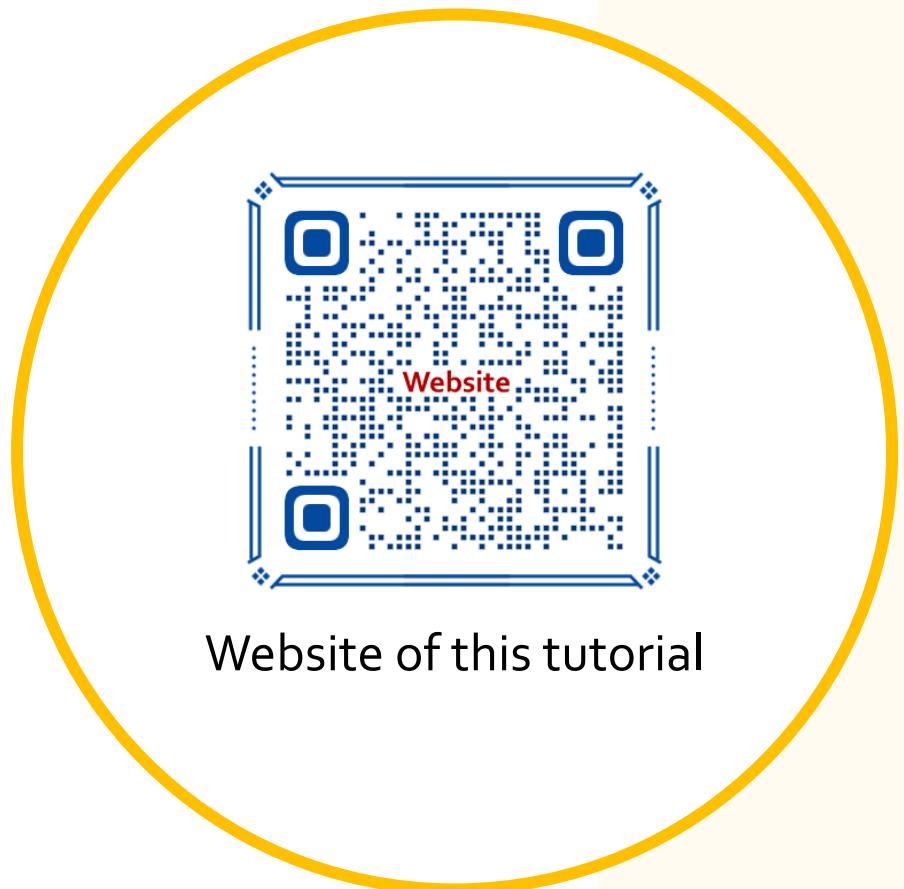
Architectures of WebAgents



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



PART 3: Architectures of WebAgents

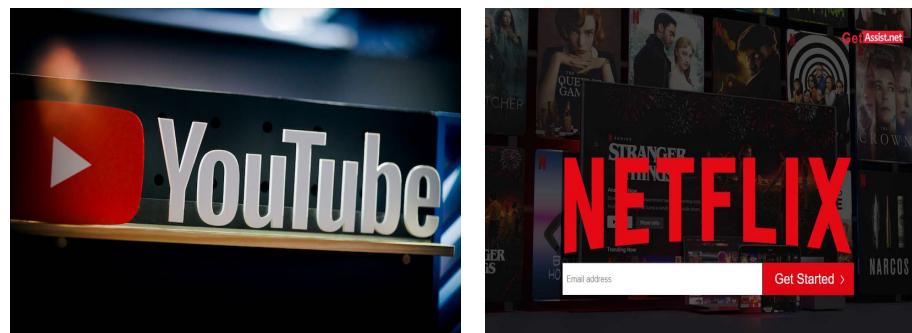
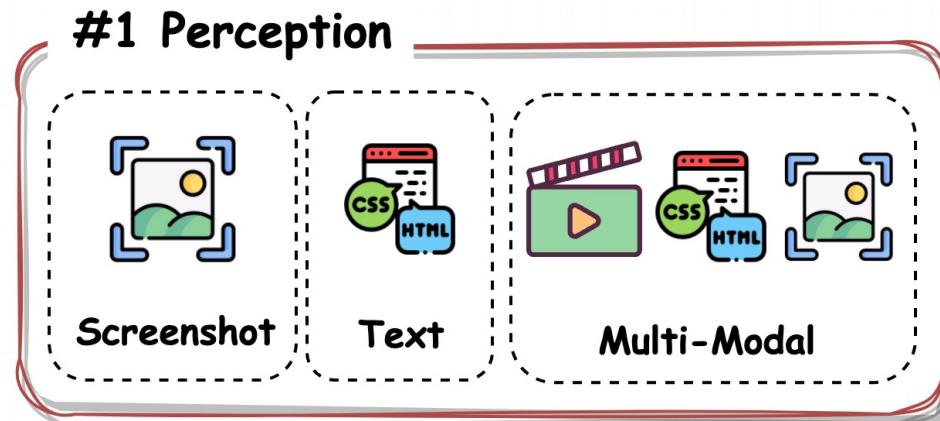


- **Perception**
- **Text-based WebAgents**
- **Screenshot-based WebAgents**
- **Multi-modal WebAgents**
- Planning & Reasoning
 - Task Planning
 - Action Reasoning
 - Memory Utilization
- Execution
 - Grounding
 - Interacting

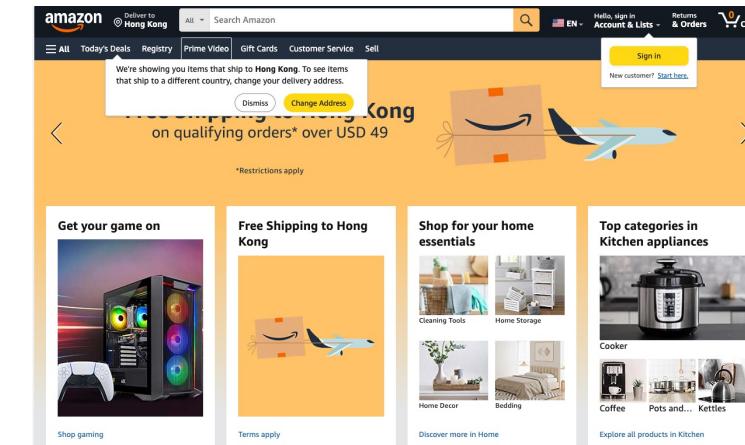
Perception



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



Video



Screenshots

[1]: Delivering to Santa Clara 95050
[2]: <input> "Search Amazon"

...

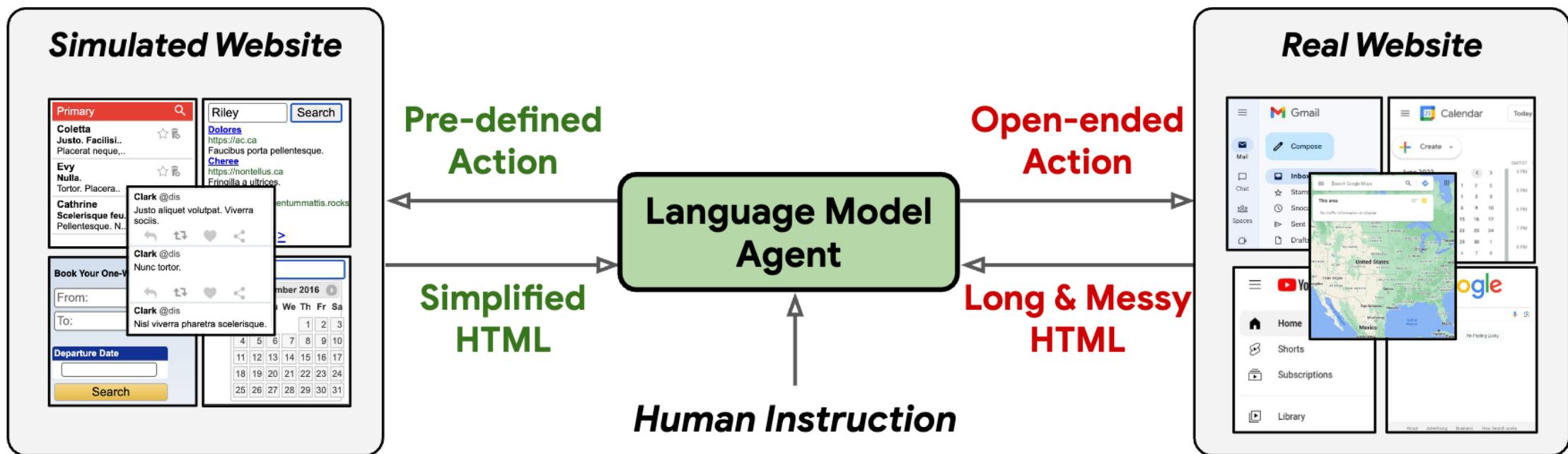
HTML

Text-based WebAgents

□ HTML-T5

Challenges in real-world web automation:

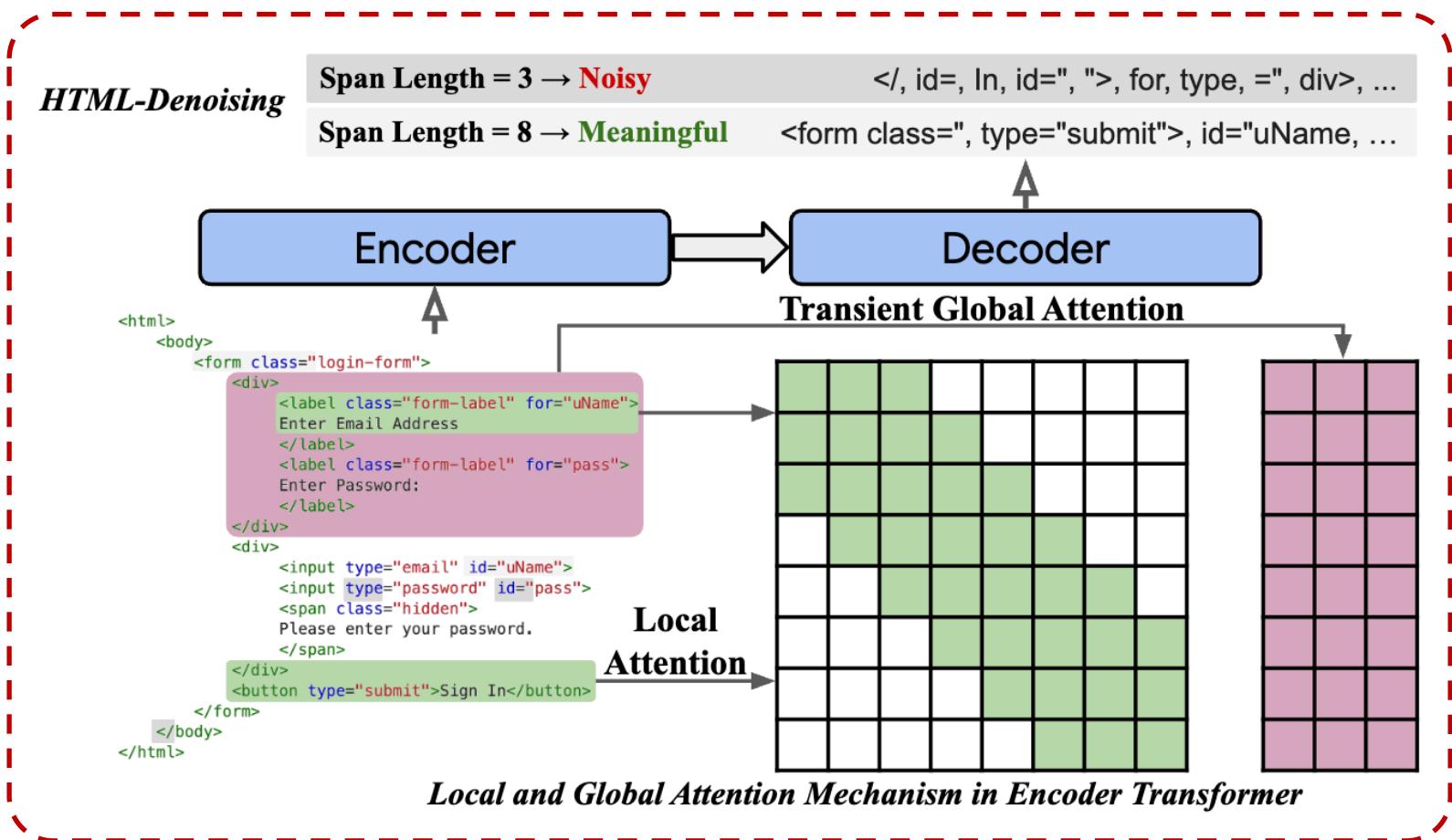
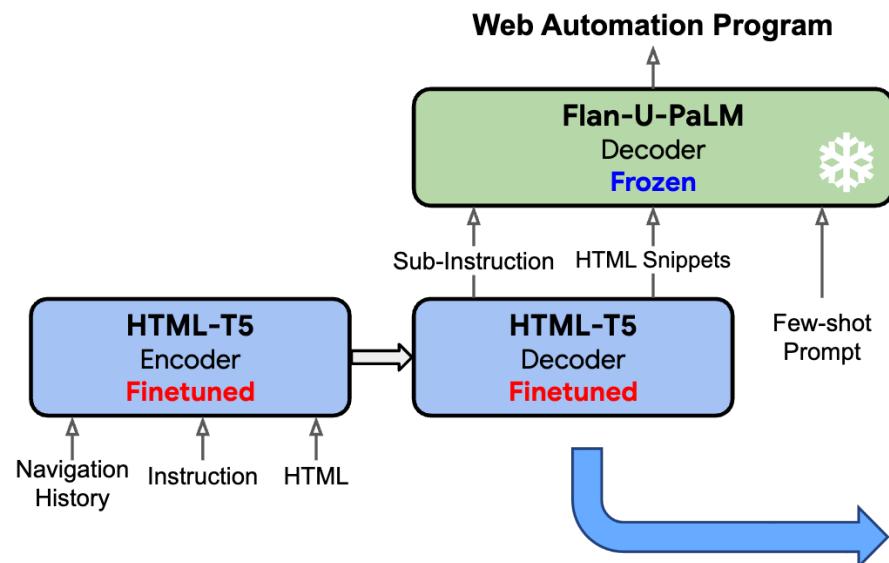
- Dynamic environments,
- open-ended actions,
- lengthy HTML documents,
- ...



Text-based WebAgents

□ HTML-T5

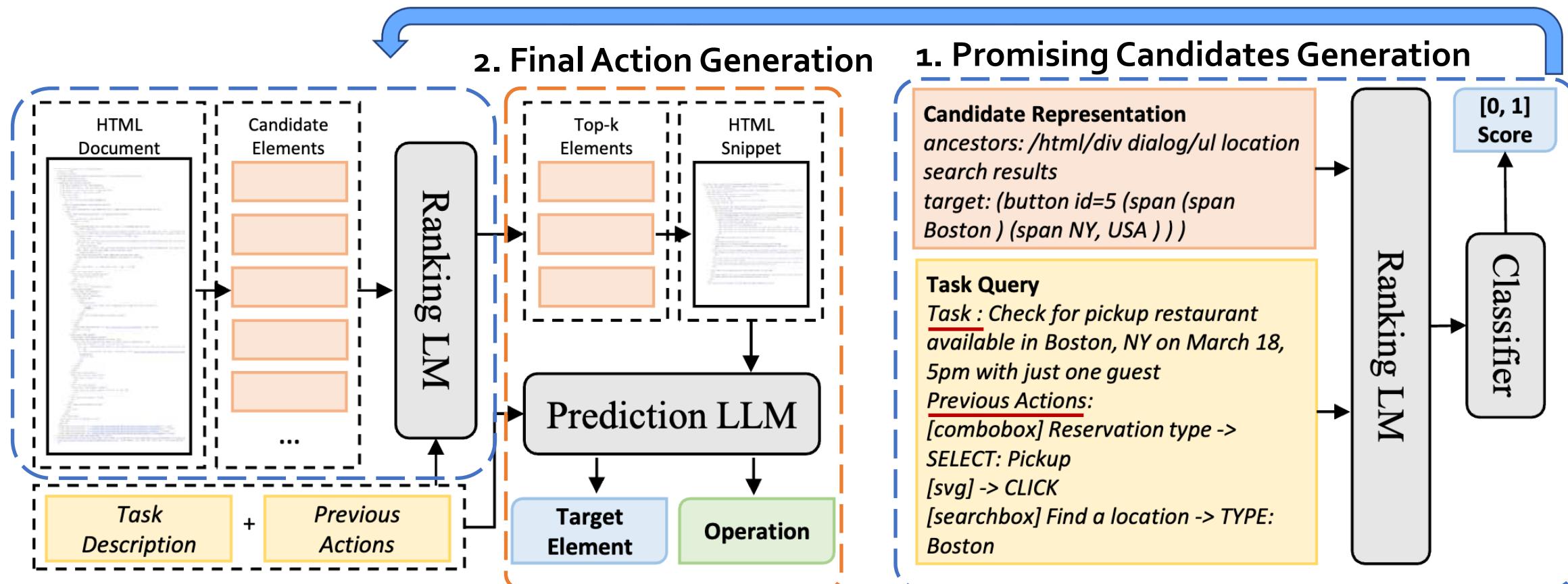
General-purpose LLMs do not fully utilize the HTML-specific information.



Text-based WebAgents



□ Mind2Web



Text-based WebAgents



□ Mind2Web

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

```
<html> <form id=0> <div meta="navigation; sitelinks">
<p> <a> Collect Renaissance </a> <a> Shop Le Meridien
</a> <a> Westin Store </a> <a> Sheraton Store </a>
</p> </div> ... <div> <select id=1 meta="Size; Select a
Size"> <span meta=tablist> <button id=2 meta="button;
tab"> Description </button> ... <a id=3 meta="Shop
Feather & Down Pillow"> <img meta="Product Feather &
Down Pillow"> <p> <a> California Privacy Rights </a>
<a> Privacy Statement </a> <a> Terms of Use </a> <a
id=4> Loyalty Terms </a> ...
```

Based on the HTML webpage above, try to complete the following task:

Task: Search for queen-size pillow protectors from the Marriott shop, and if found, add two pieces to the cart and checkout.

Previous actions:

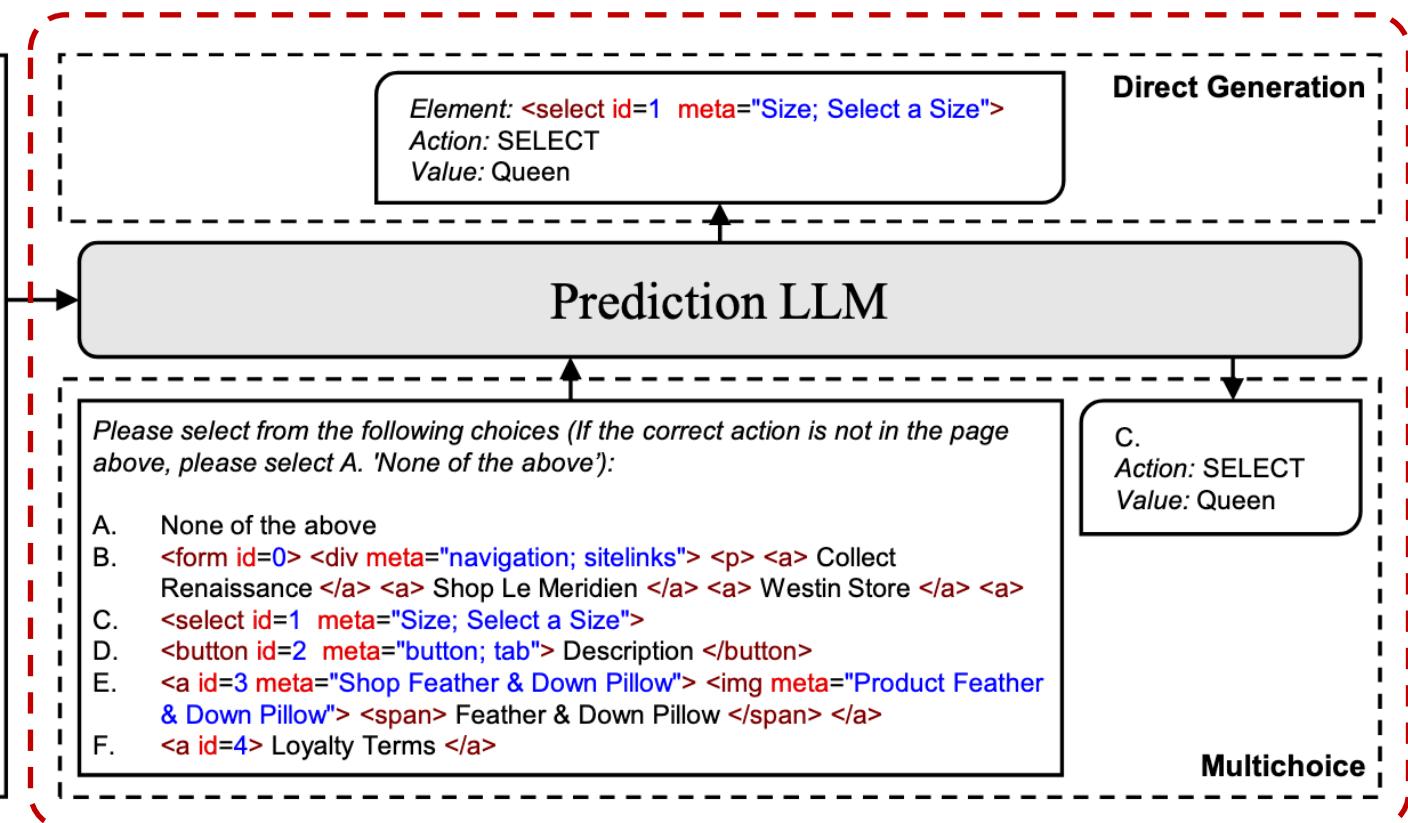
[button] Special Offers -> CLICK

[link] Shop Marriott Opens a new window -> CLICK

[menuitem] category pillows -> CLICK

[span] Pillow Protector -> CLICK

What should be the next action?



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:

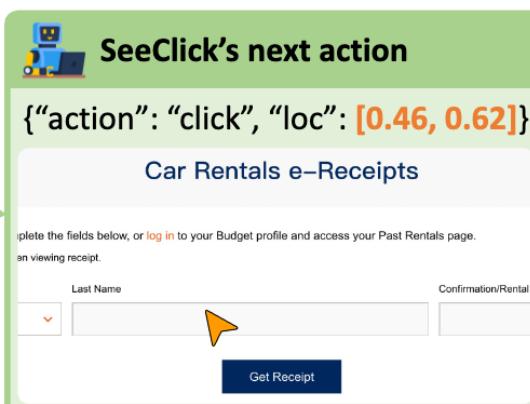
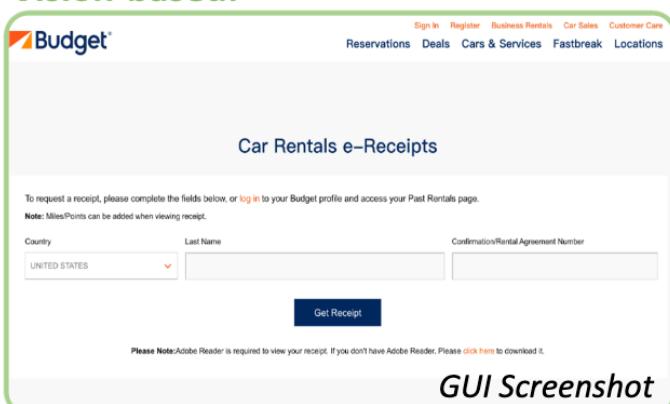
```
<form element_id="200">  
...  
    <label element_id="205">Last Name:</label>  
    <input type="text" name="lastname" element  
    _id="206">  
...  
    <input type="submit" value="Get Receipt" element  
    _id="210">  
...
```

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:

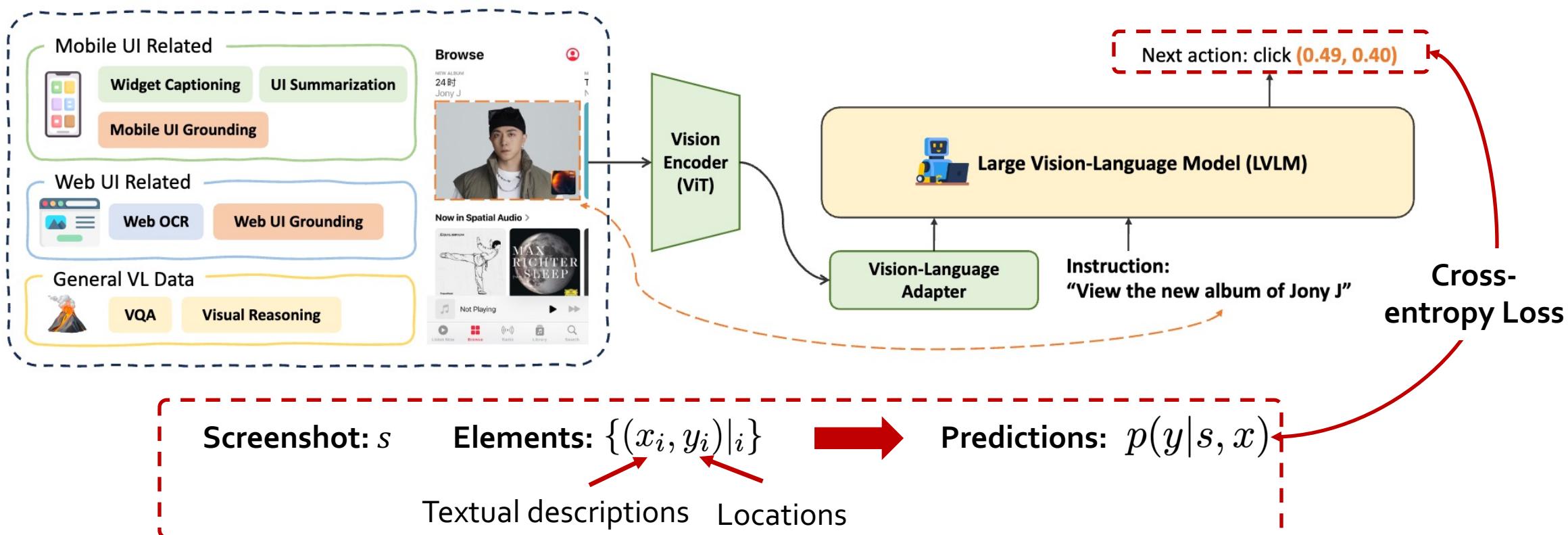


Screenshot-based WebAgents



□ SeeClick

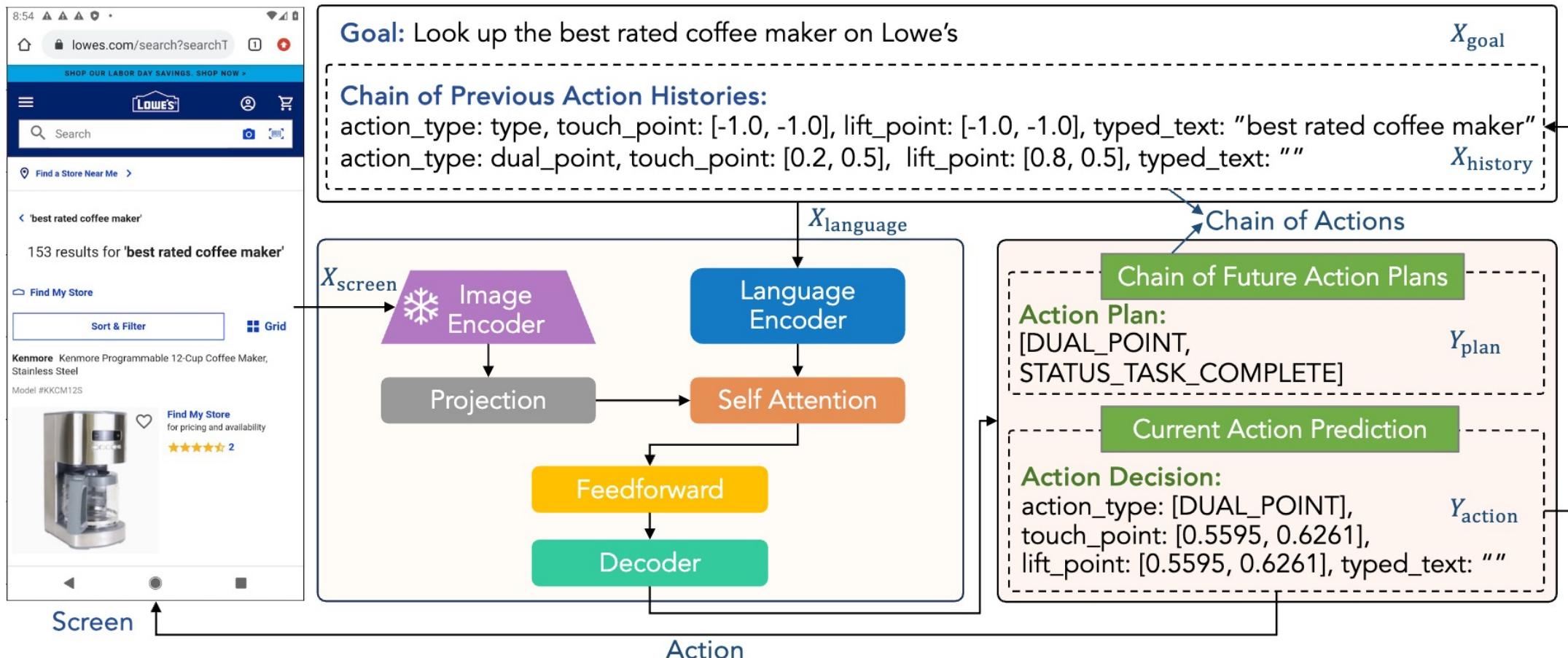
GUI grounding - the capacity to accurately locate screen elements based on instructions, which is absent in current LVLMs.



Screenshot-based WebAgents



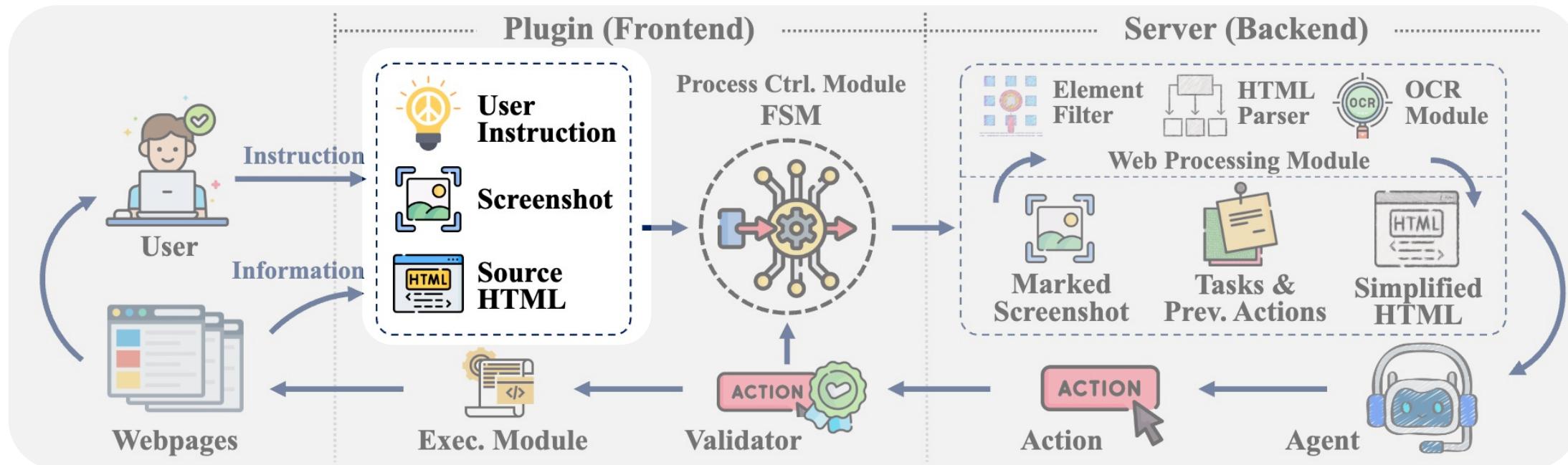
□ Auto-GUI



Multi-Modal WebAgents



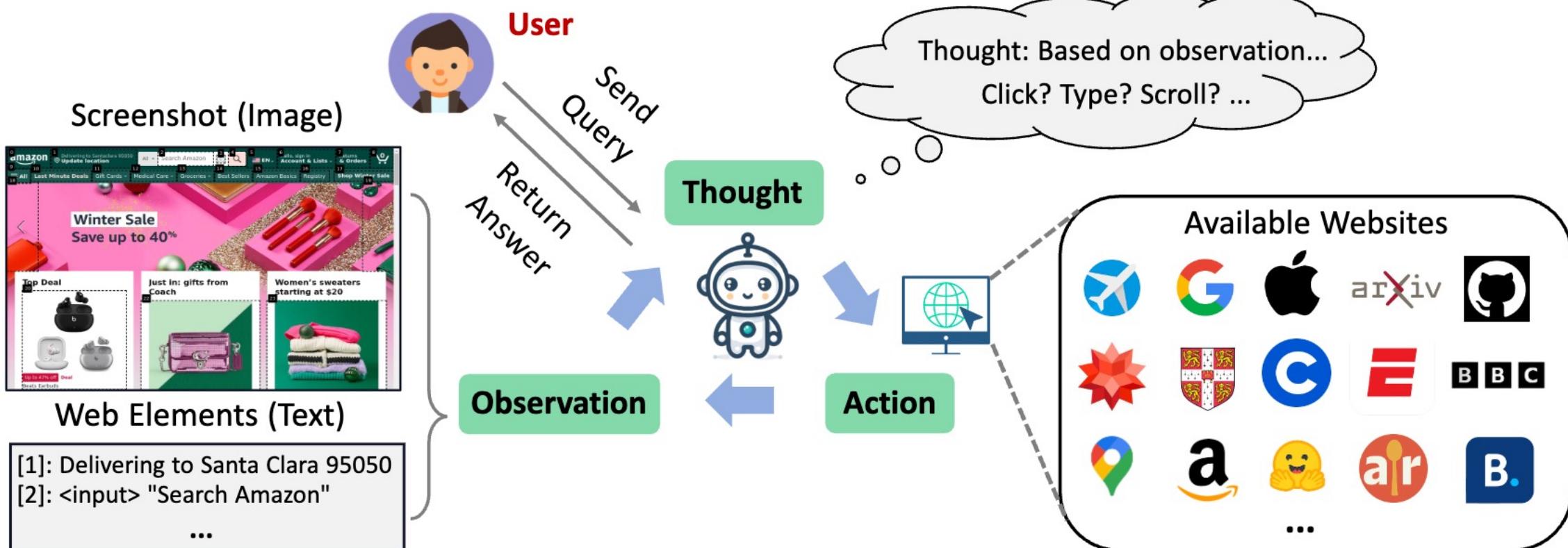
- ❑ Leveraging multi-modal data, combining their complementary strengths, can provide WebAgents with a more comprehensive environmental perception.



Multi-Modal WebAgents



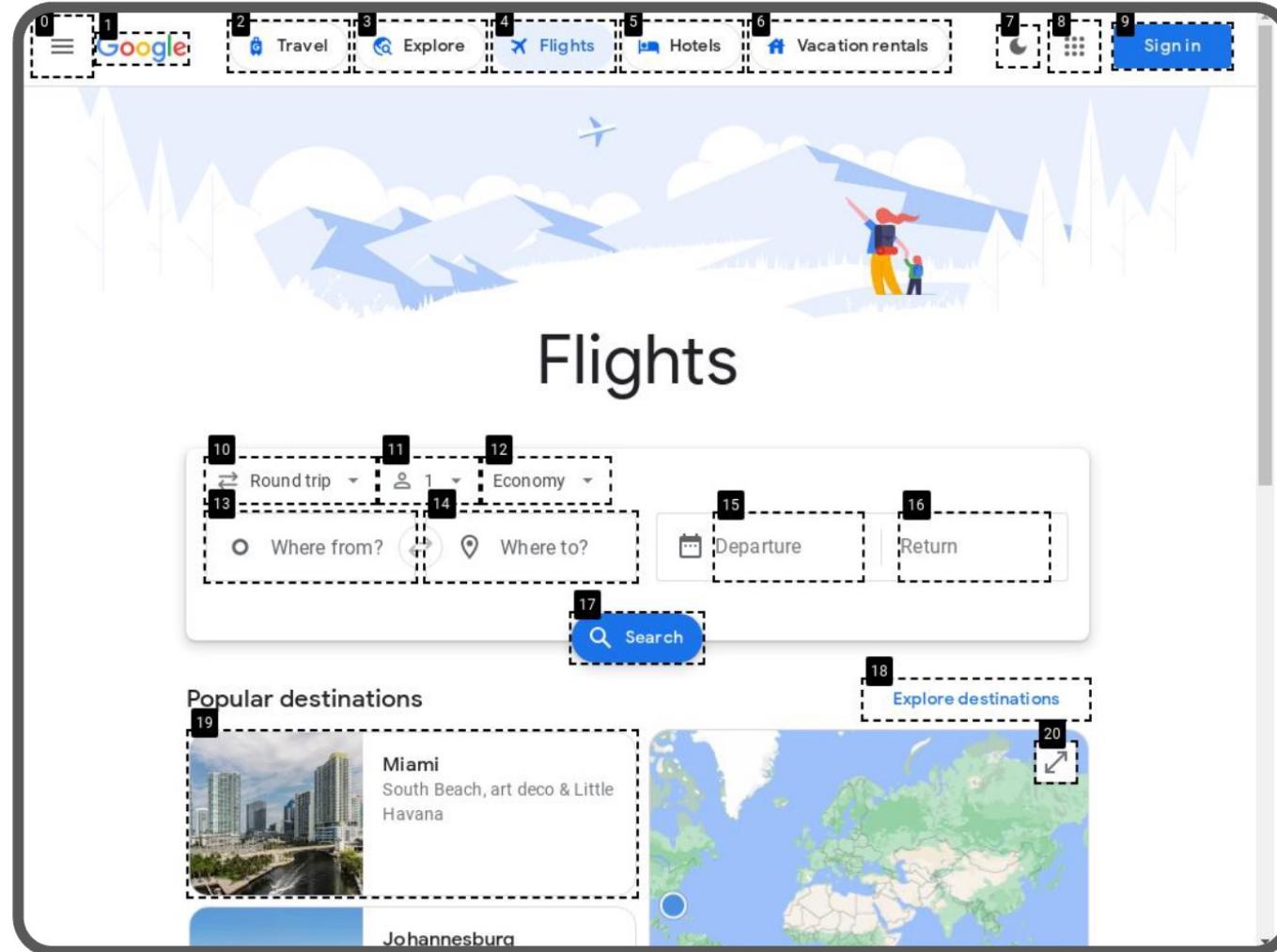
□ WebVoyager



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
 - ...

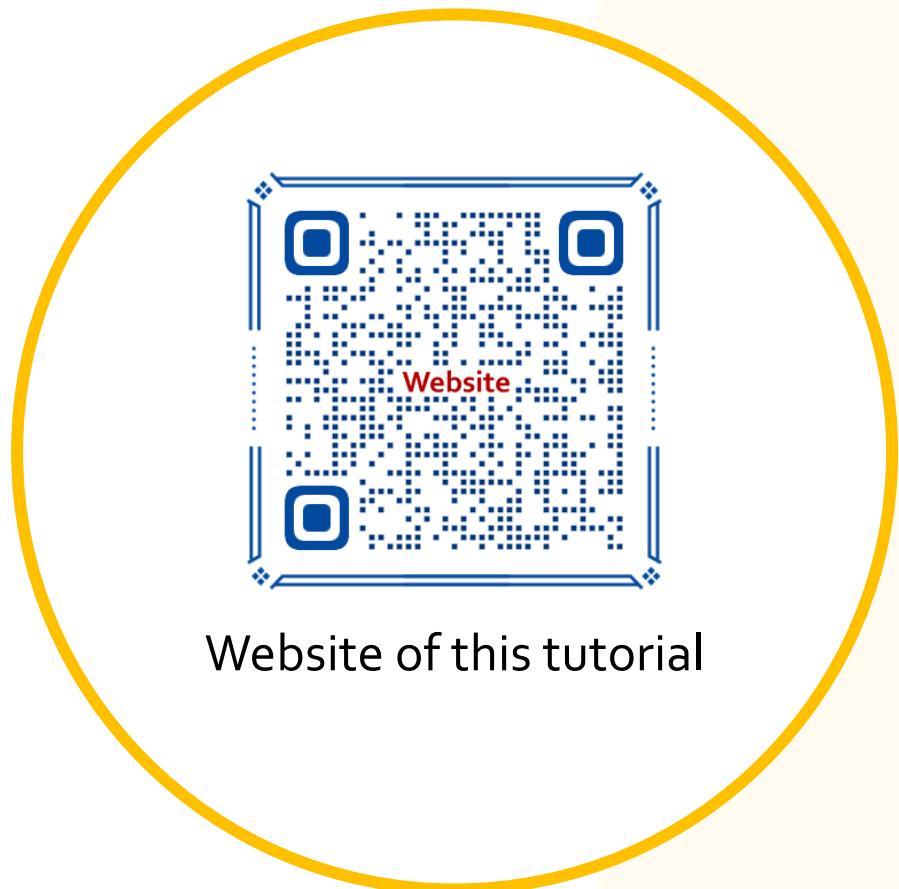
Multi-modal WebAgents

- ➤ Comprehensive information
 - ➤ Robust
 - ...
- ➤ High resource use
 - Complex pipelines
 - Potential redundancy
 - ...



More Popular

PART 3: Architectures of WebAgents

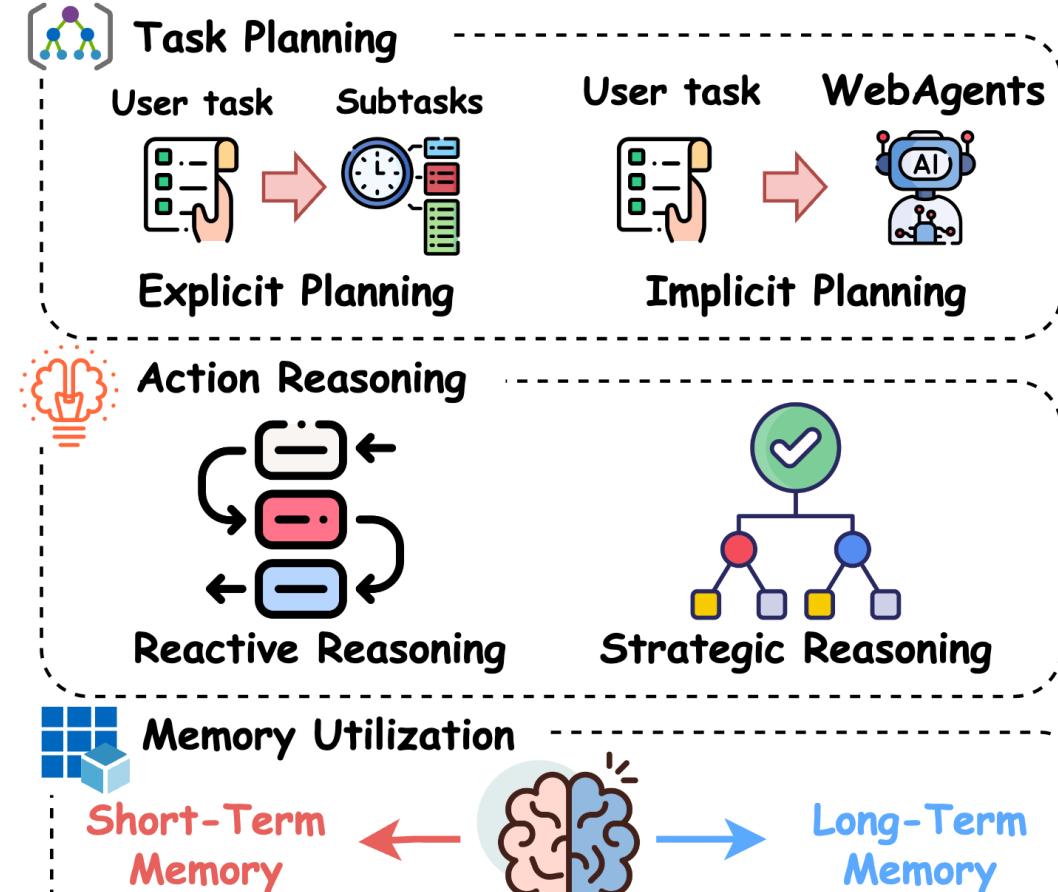


- Perception
- Text-based WebAgents
- Screenshot-based WebAgents
- Multi-modal WebAgents
- **Planning & Reasoning**
- Task Planning
- Action Reasoning
- Memory Utilization
- Execution
- Grounding
- Interacting

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.

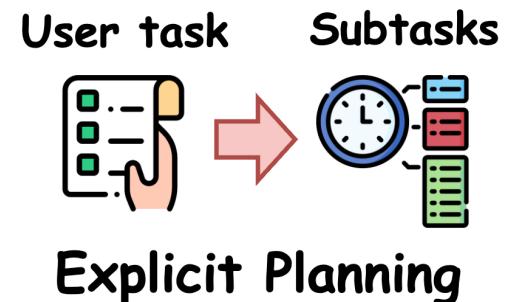
#2 Planning & Reasoning



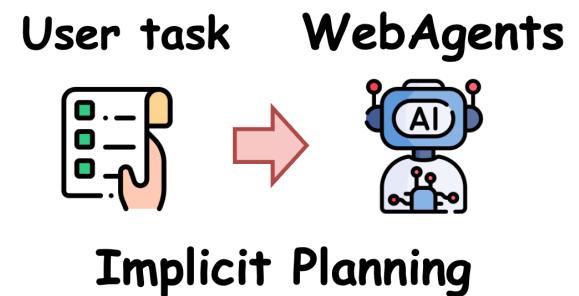
Task Planning

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.

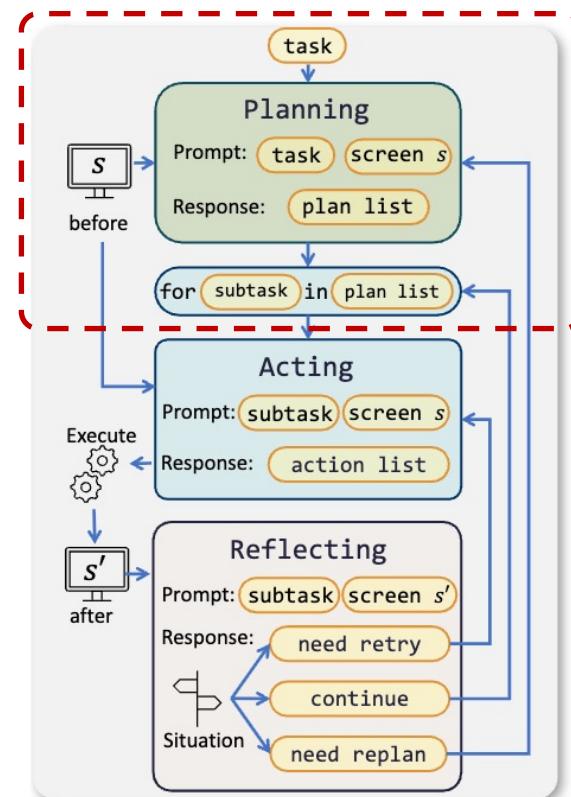


Task Planning – Explicit Planning

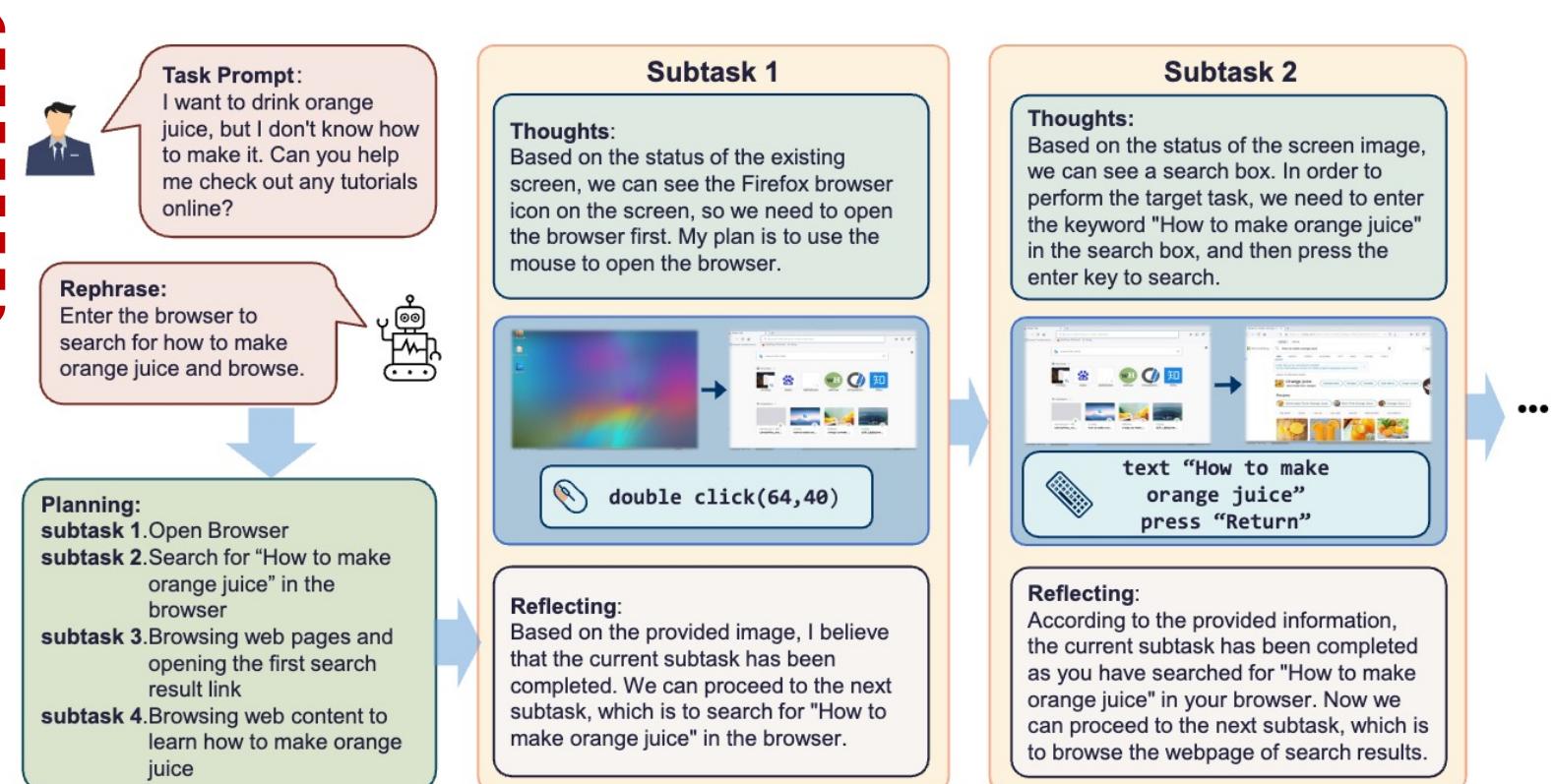


ScreenAgent

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



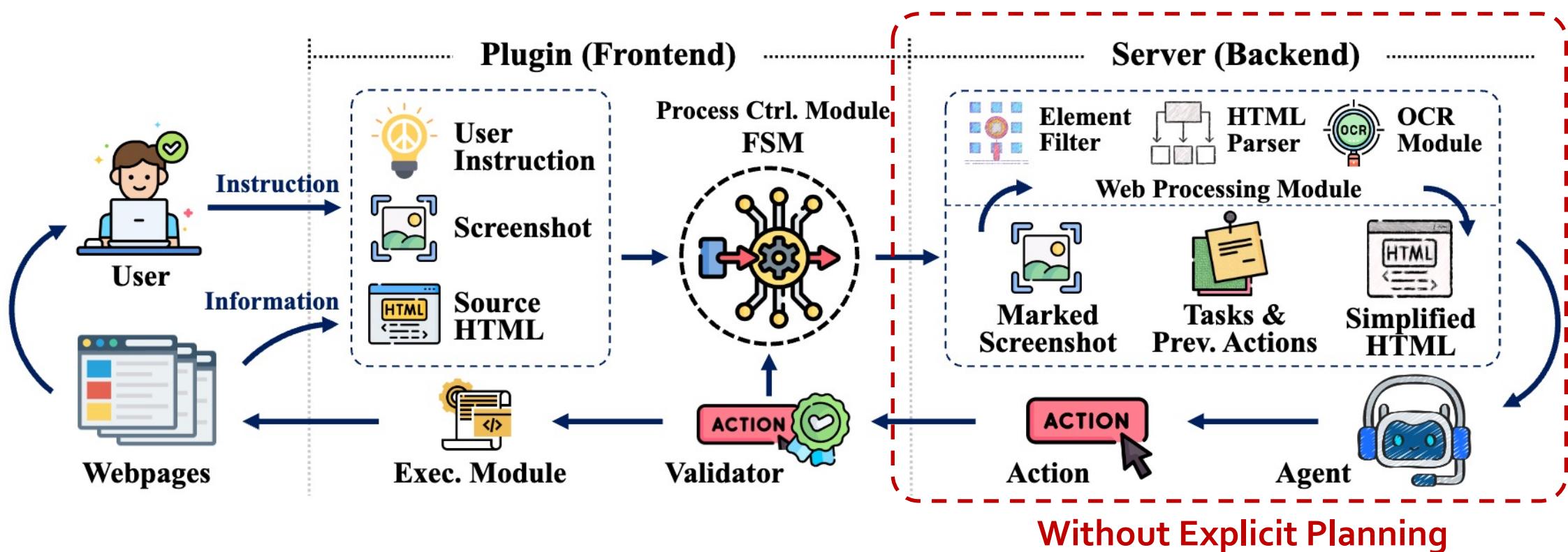
(a) Flowchart of our pipeline



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
- Black-box process
- ...

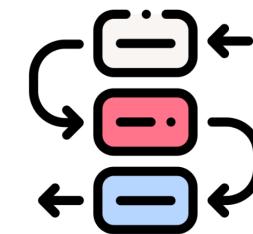


Action Reasoning

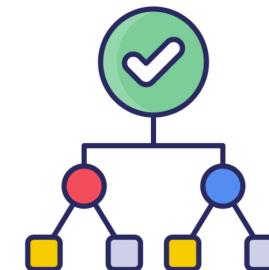


Action reasoning involves leveraging the agent's reasoning capabilities and current environmental observations to **infer the next action**.

- **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.



Reactive Reasoning

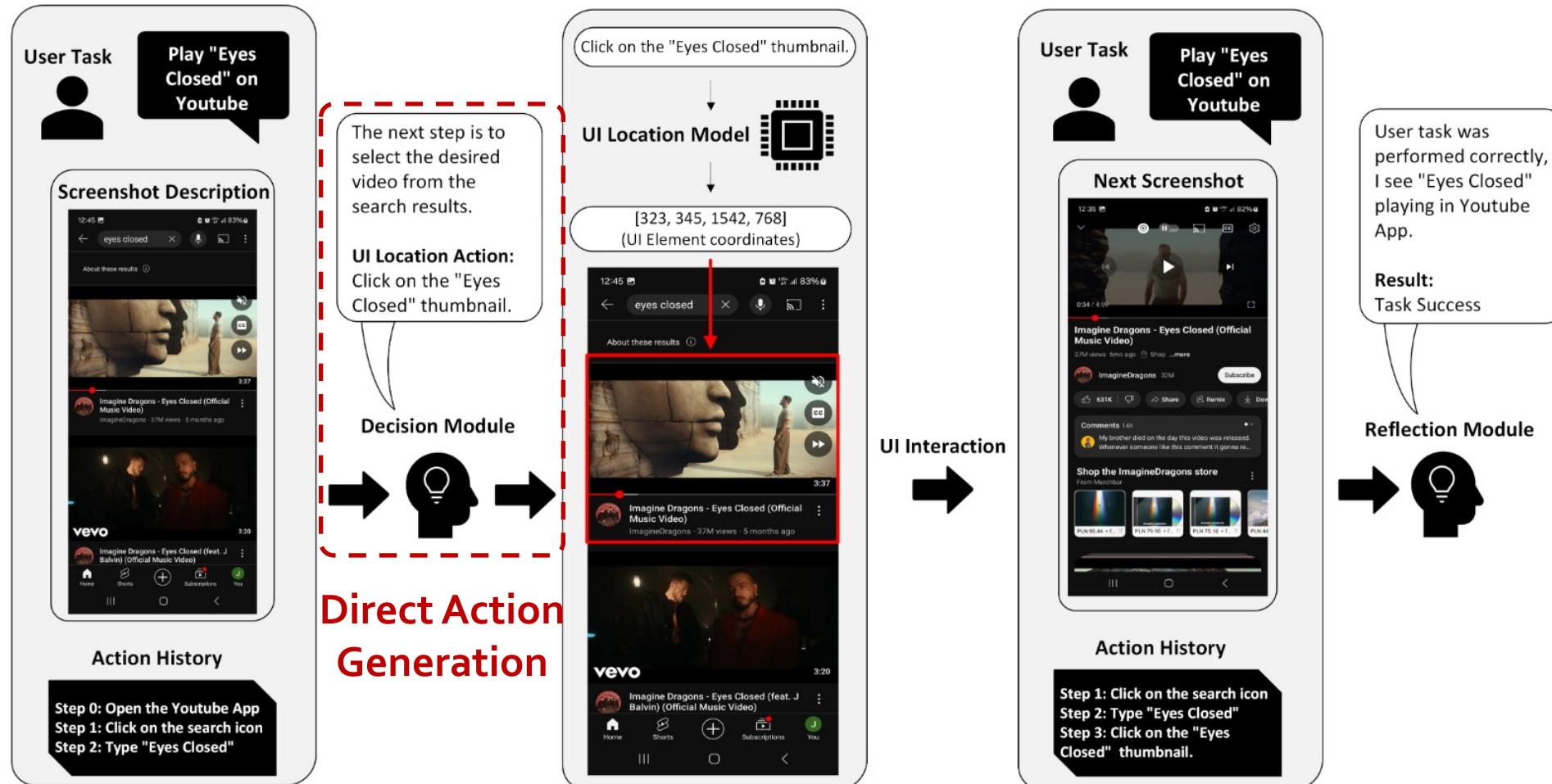


Strategic Reasoning

Action Reasoning – Reactive Reasoning



□ ClickAgent

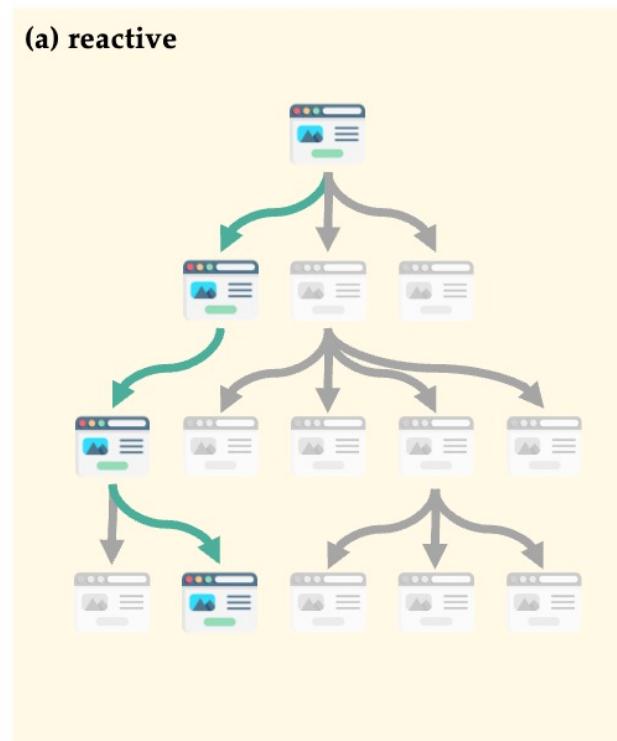


Action Reasoning – Strategic Reasoning

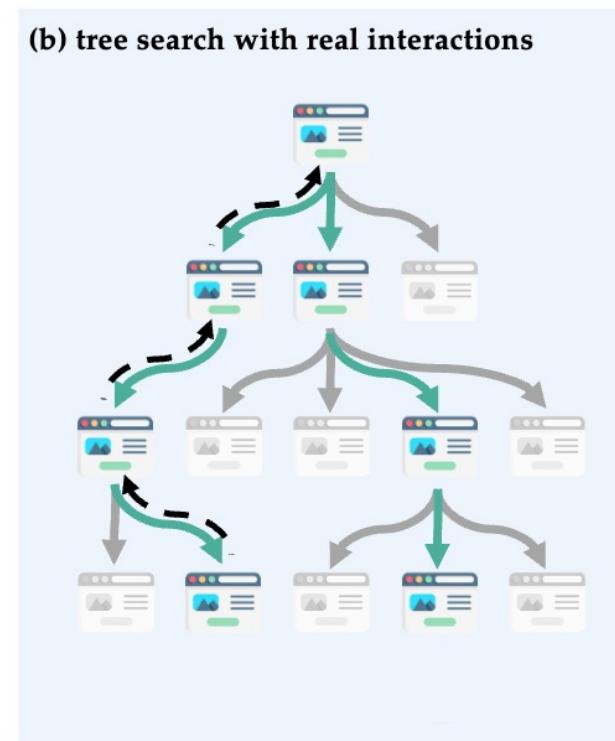


 WebDreamer

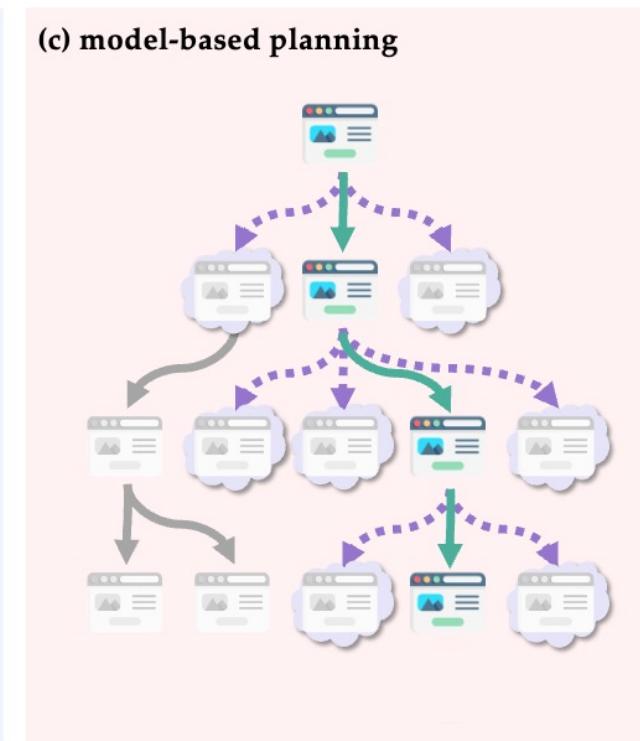
LLM-driven exploration strategy to simulate and predict the outcomes of candidate actions before execution using natural language descriptions.



Locally optimal actions



Backtracking is infeasible

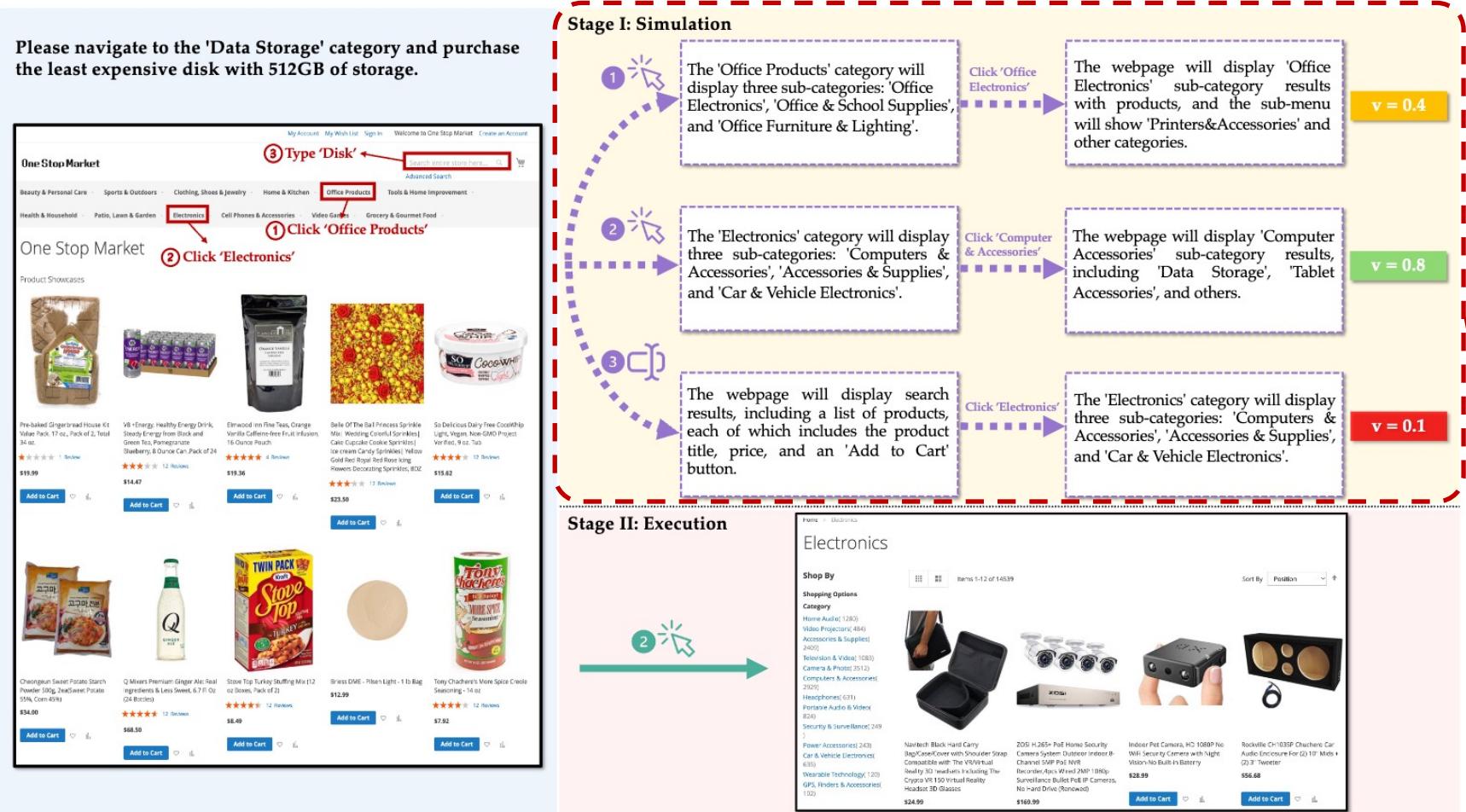


Backtracking is infeasible

Action Reasoning – Strategic Reasoning



□ WebDreamer



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**

```

 $A_t \leftarrow \text{get\_candidate}(I, o_t);$ 
 $A'_t \leftarrow \text{self\_refine}(A_t);$ 
 $a_t = \arg \max_{a \in A'_t} [\text{score}(\text{sim}(o_t, a))];$ 
 $o_{t+1} \leftarrow \text{execute}(a_t);$ 
 $t \leftarrow t + 1;$ 

```

```

if termination_check() = True then
| break;
end

```

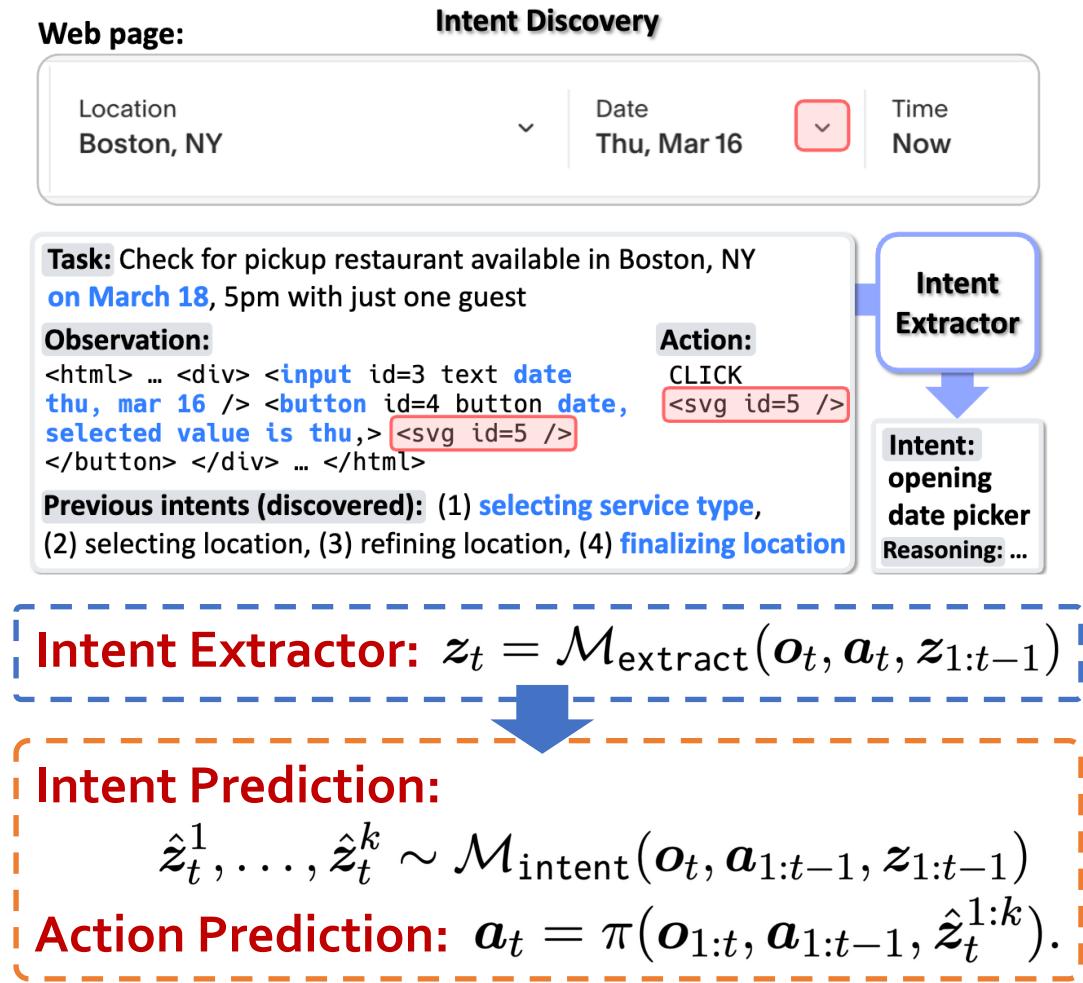
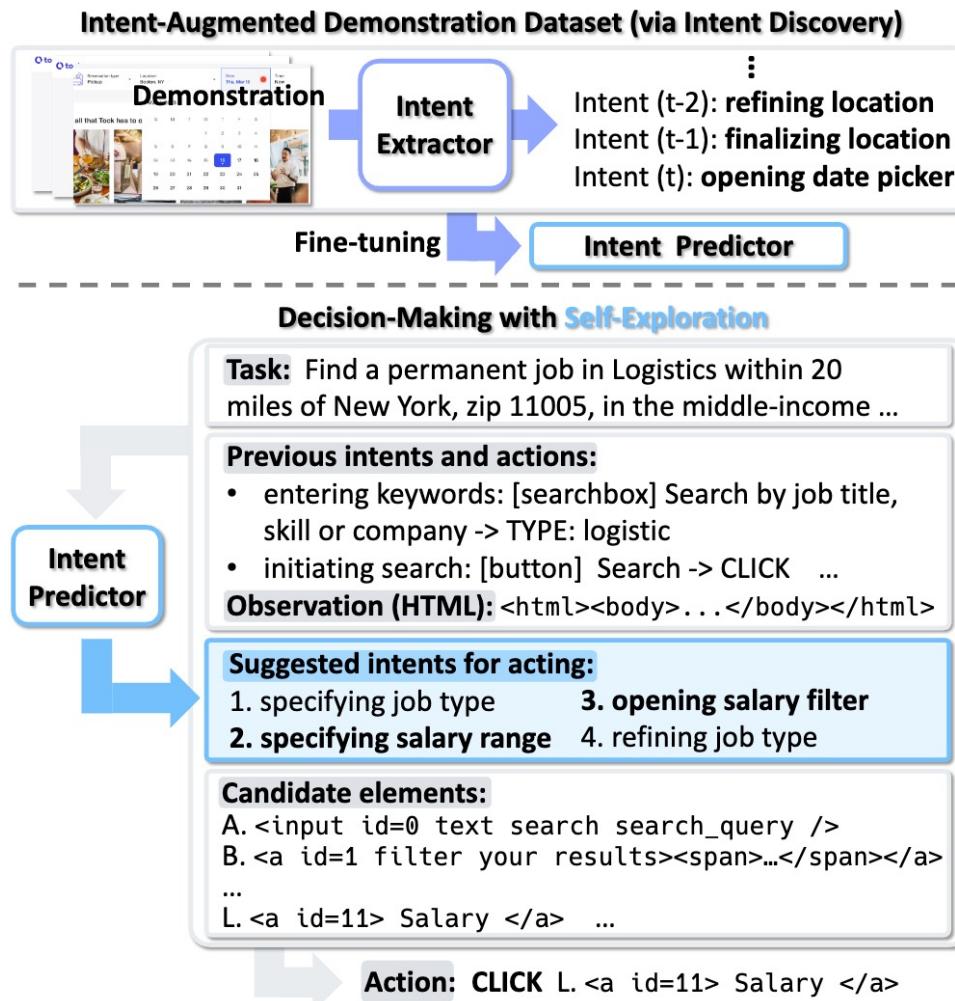
end

Return result;

Action Reasoning – Strategic Reasoning



Auto-Intent



Reasoning



❑ Reactive Planning

- Efficient
- Easy to implement
- ...



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



❑ Strategic Reasoning

- More accurate action prediction
- ...

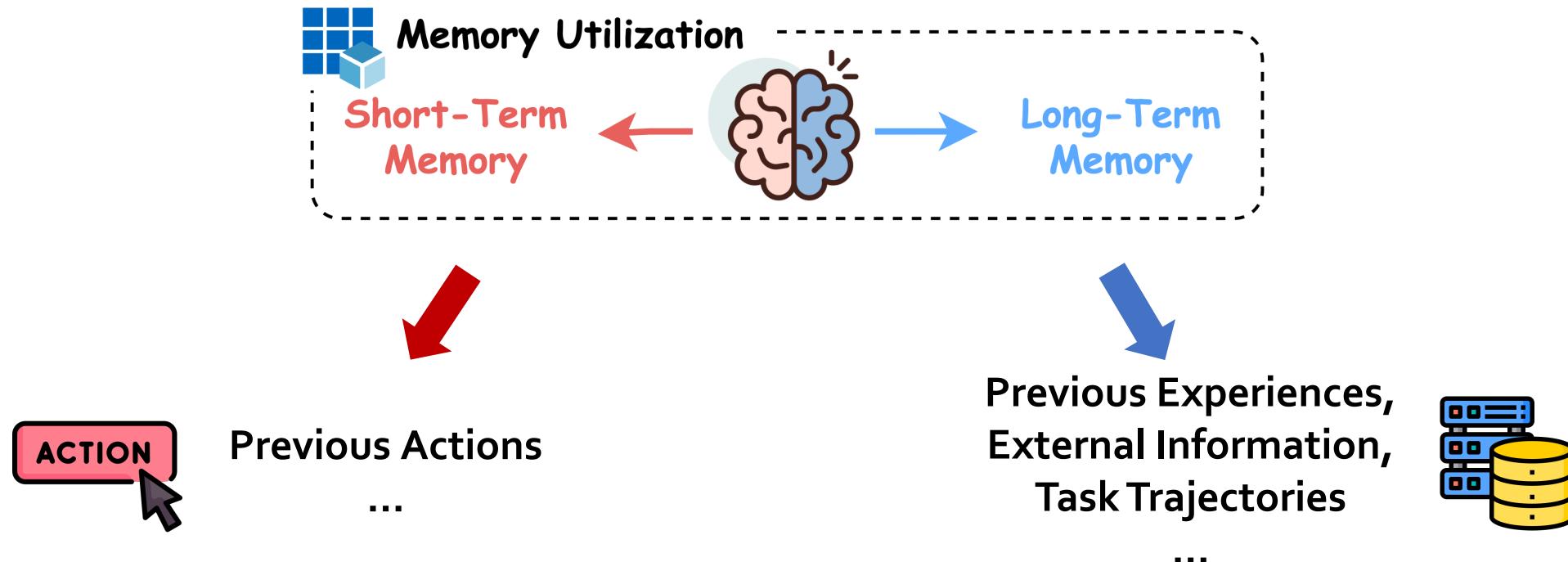


- Additional computational cost
- Design sophisticated mechanisms.
- ...

Memory Utilization

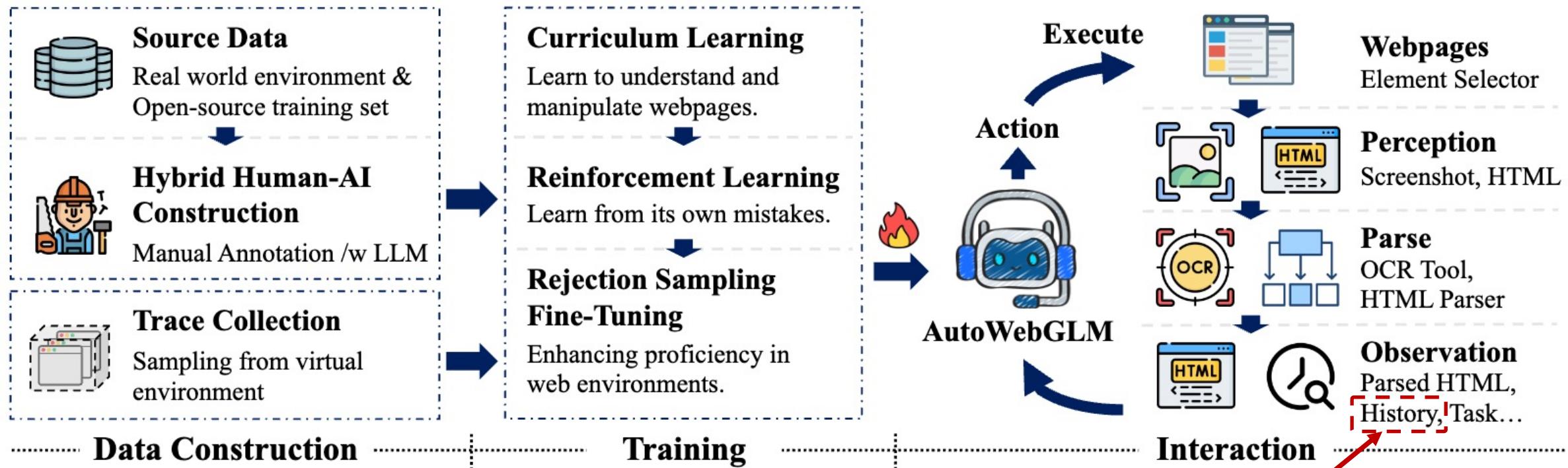


- ❑ Beyond task planning and action reasoning, **the effective utilization of memory is another key factor contributing to the powerful capabilities of WebAgents.**



Memory Utilization – Short-Term

AutoWebGLM



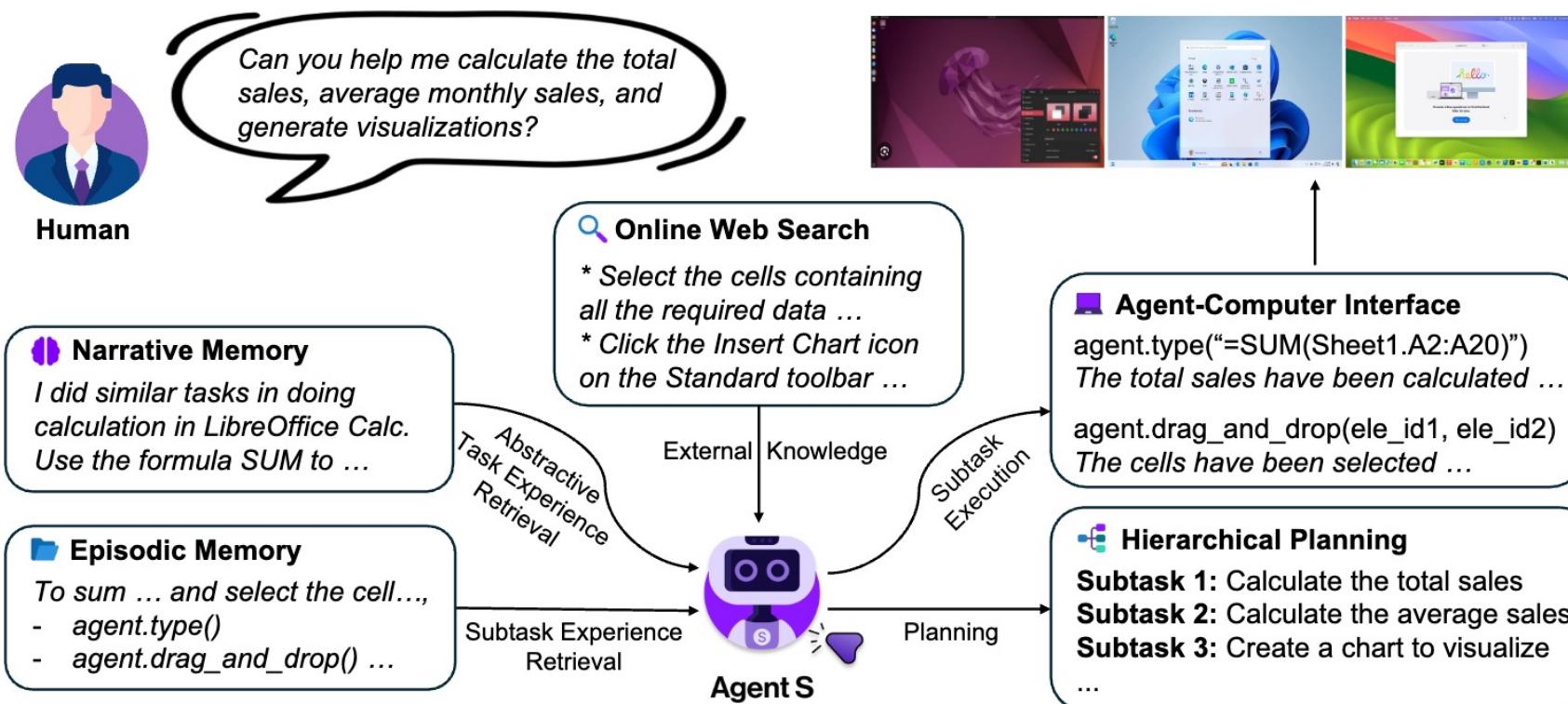
The best solution to inform the agent of past operations is explicitly providing it.

Memory Utilization – Long-Term



❑ AGENTS

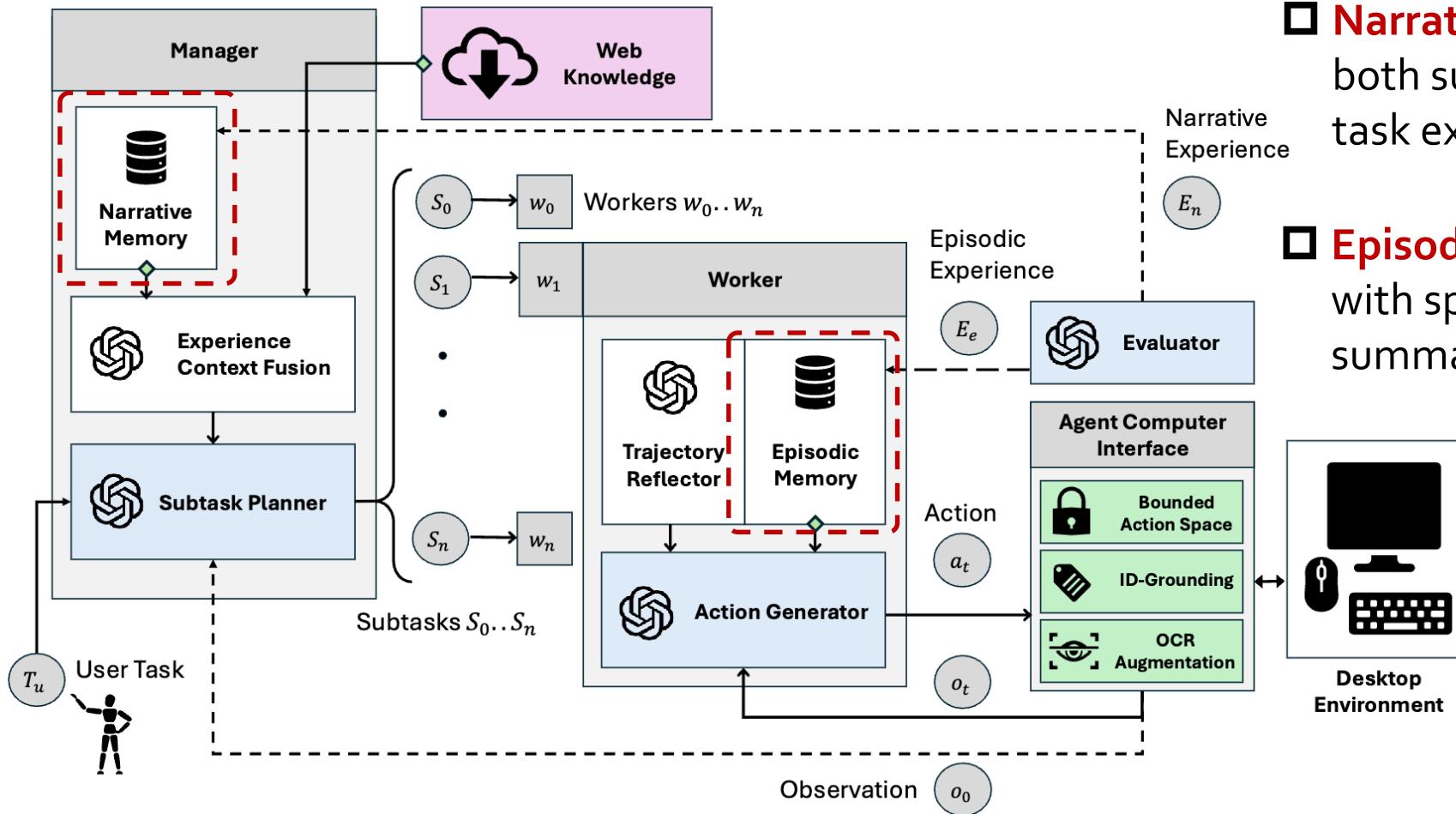
- Leverages **Online Web Knowledge** and past experiences stored in **Narrative Memory** to decompose the complex, long-horizon task into a structured plan of manageable subtasks
- Retrieves step-by-step subtask experience from **Episodic Memory** to refine the actions.



Memory Utilization – Long-Term



❑ AGENTS



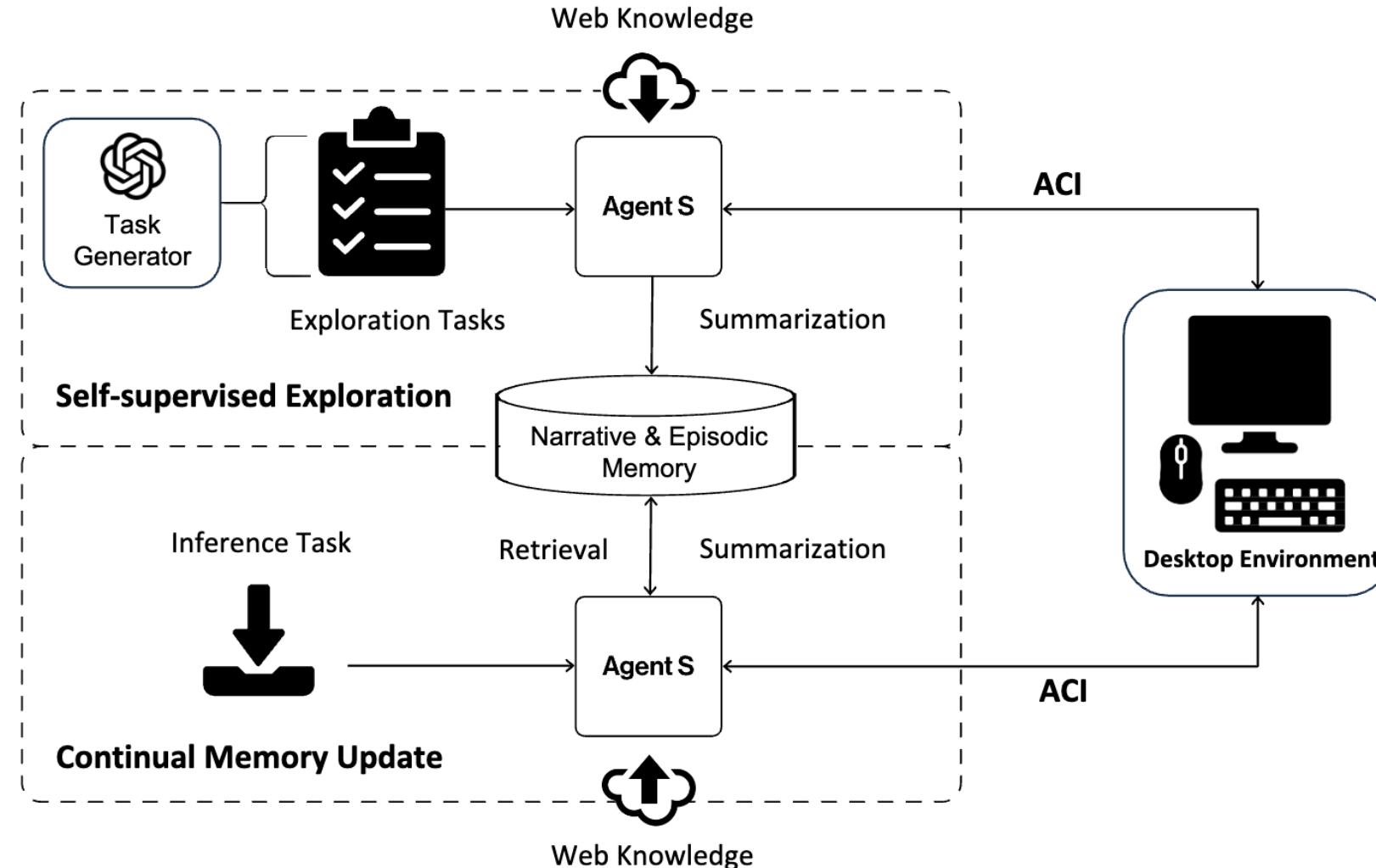
❑ **Narrative Memory:** includes summaries of both successful and failed trajectories as task experiences.

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

Memory Utilization – Long-Term



❑ AGENTS



Memory Utilization



☐ Short-term Memory

- Efficient
- Short context
- ...

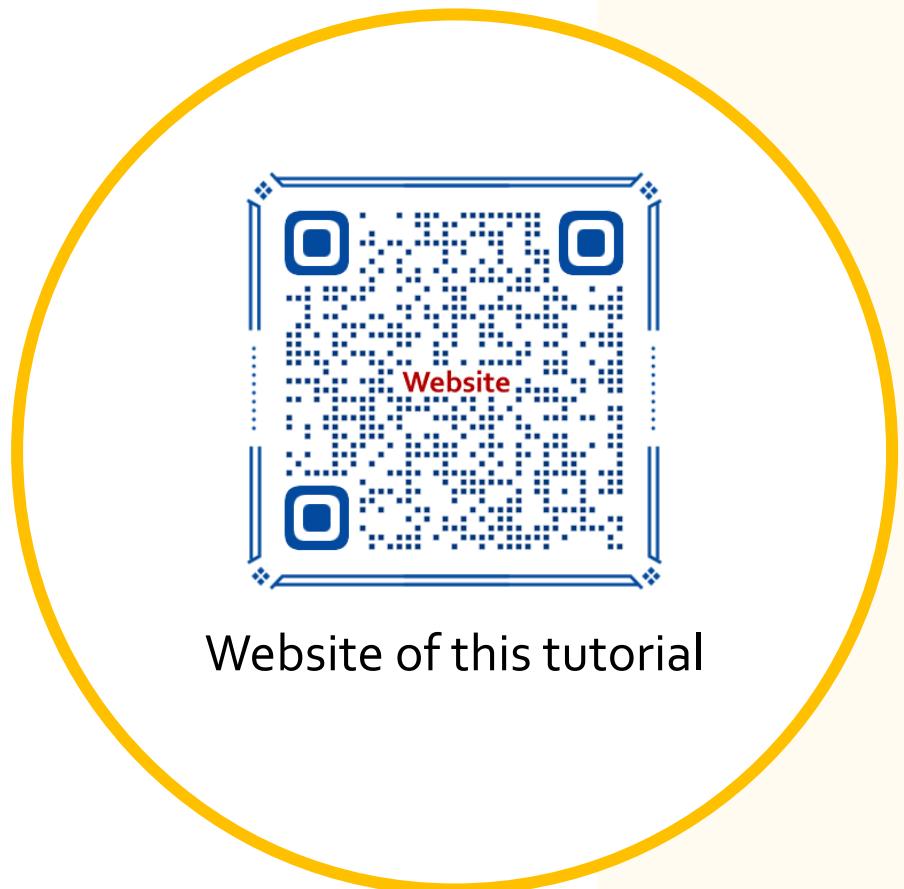
- ✖➤ Limited information
- ✖➤ Fixed knowledge
- ✖➤ ...

☐ 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

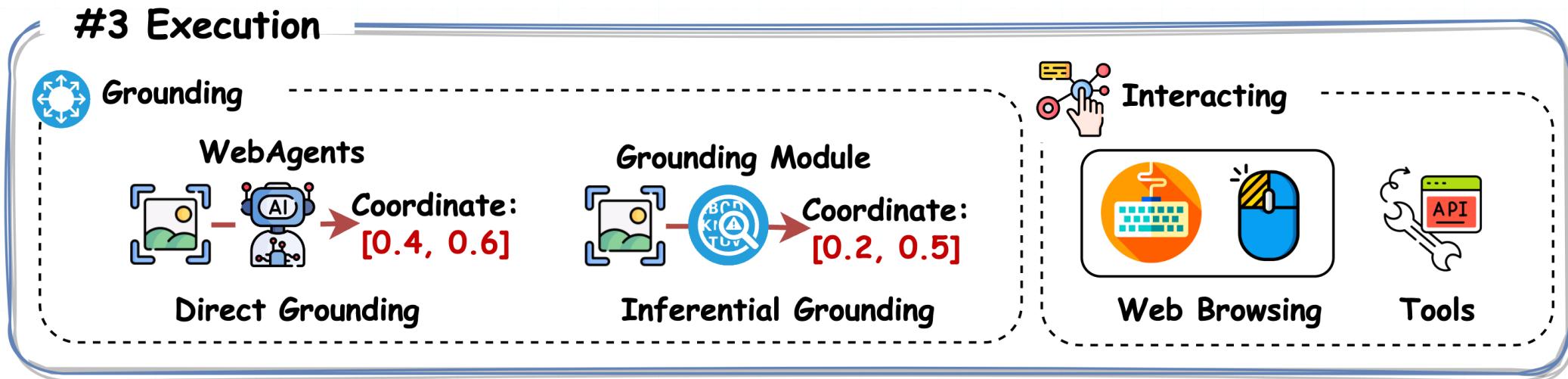


- Perception
 - Text-based WebAgents
 - Screenshot-based WebAgents
 - Multi-modal WebAgents
- Planning & Reasoning
 - Task Planning
 - Action Reasoning
 - Memory Utilization
- **Execution**
 - **Grounding**
 - **Interacting**

Execution

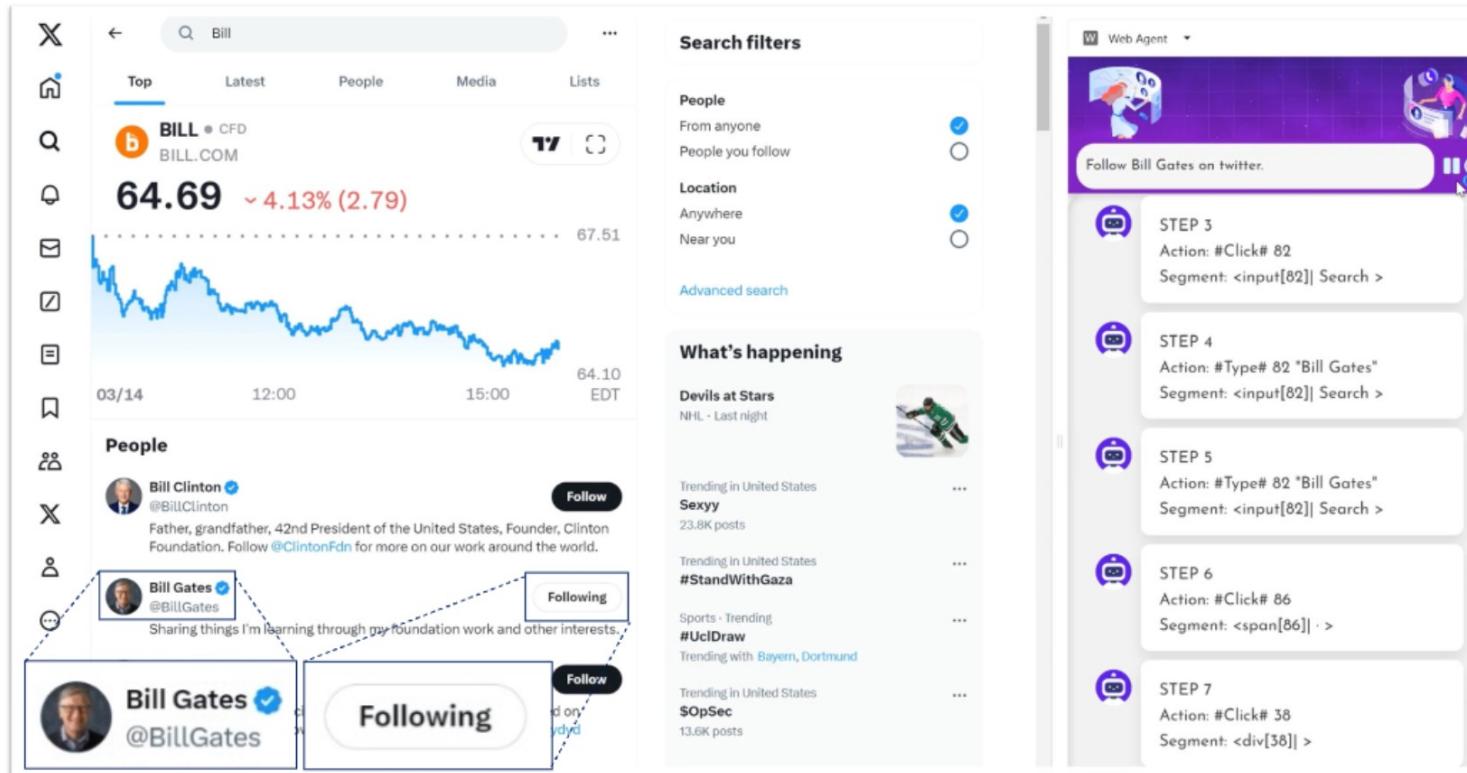


- The final step for WebAgents to complete the user's command is to **interact with the webpages and execute the generated actions.**

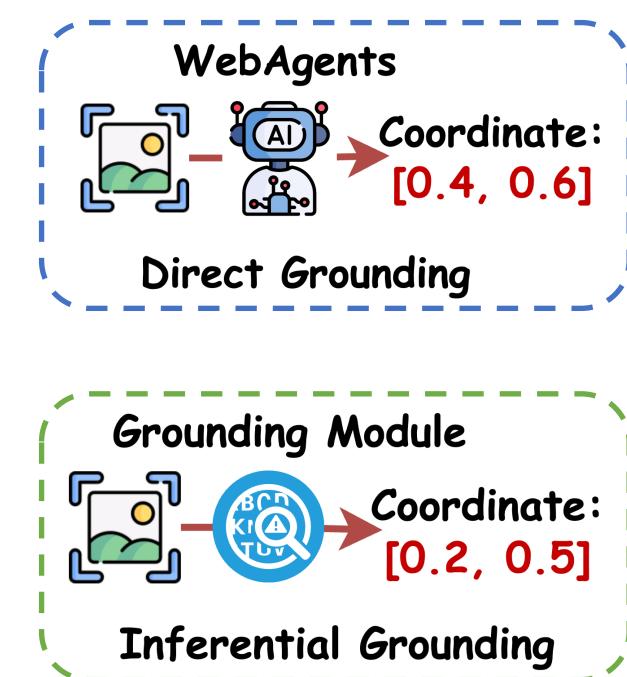


Grounding

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



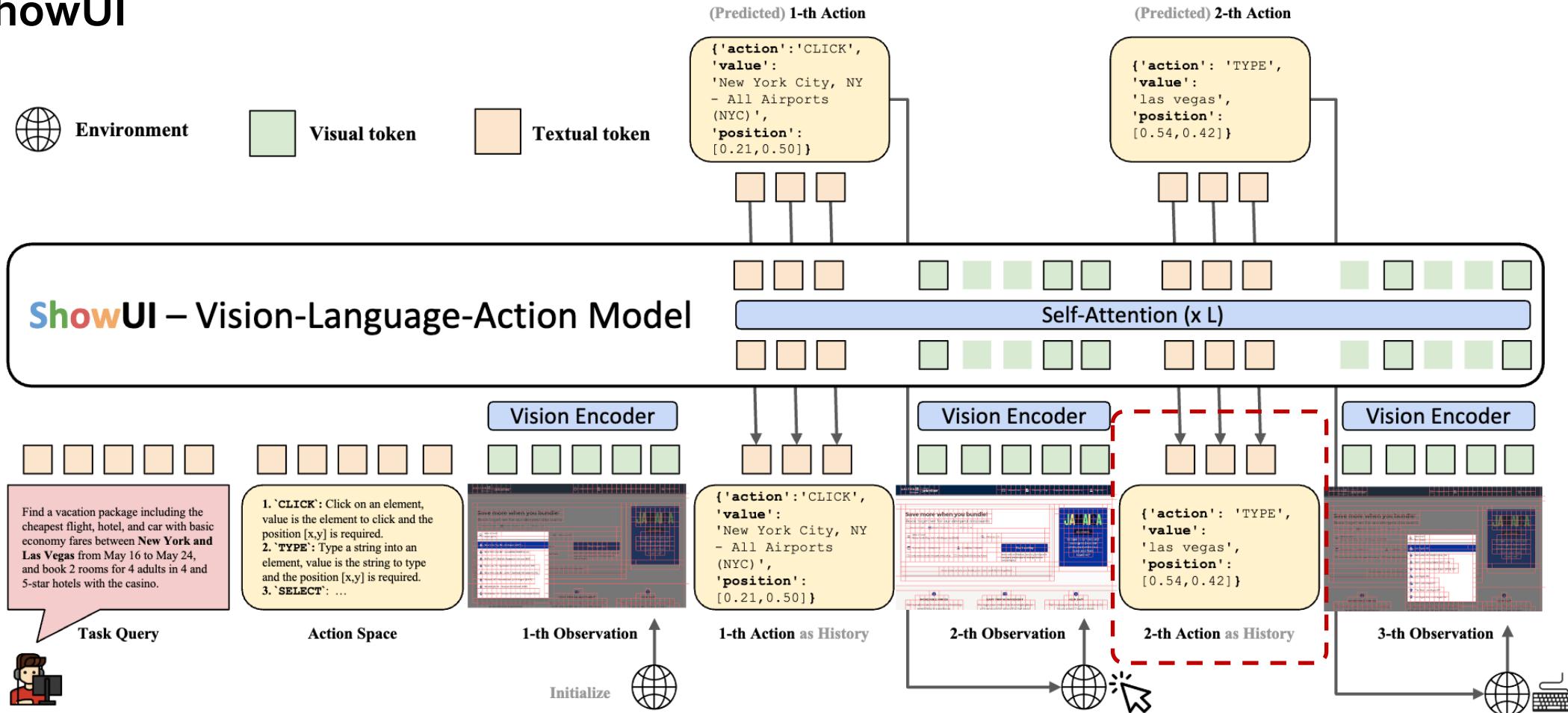
(a) Follow Bill Gates on X (Twitter).



Grounding – Direct Grounding



□ ShowUI



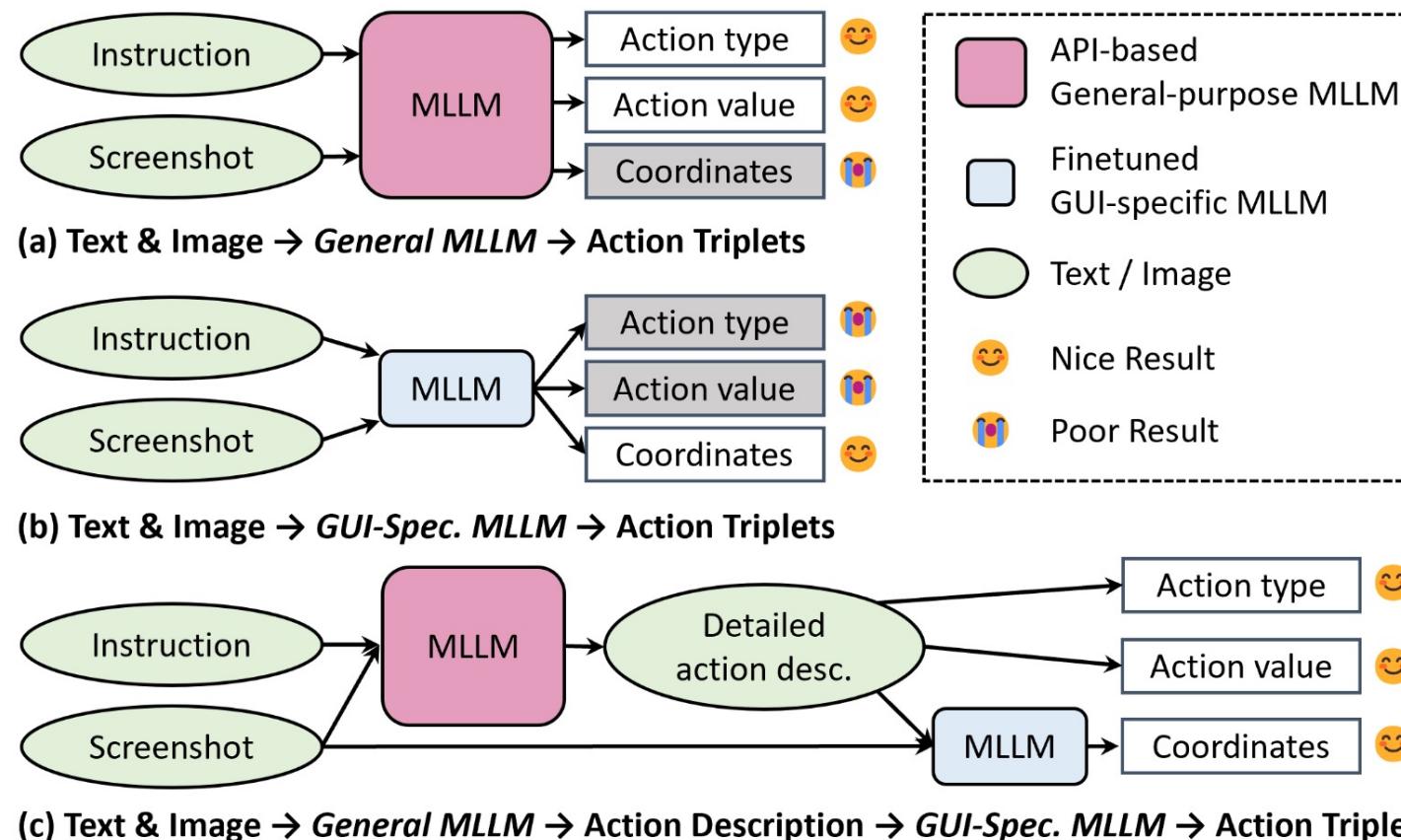
Contains both generated actions and the target element position

Grounding – Inferential Grounding



❑ Ponder & Press

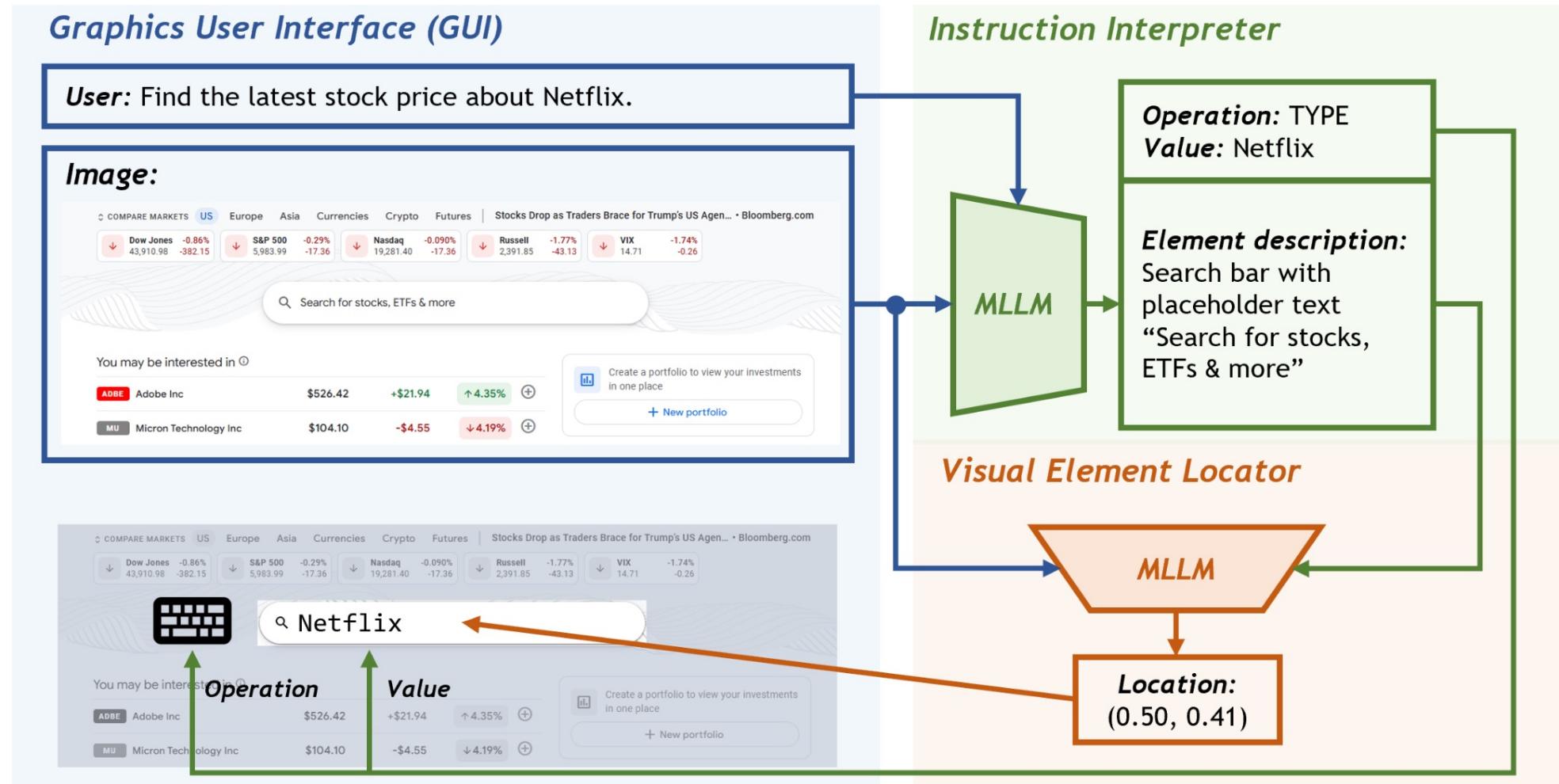
User instructions are directly mapped to action triplets in a single inference, which is **difficult due to the significant difference between the textual nature of actions and values, and the numerical nature of pixel coordinates.**



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
- ...



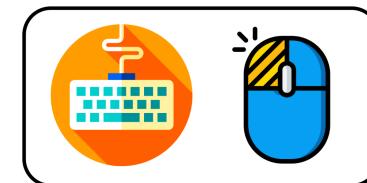
- Sophisticated mechanisms to locate the correct element
- ...

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.

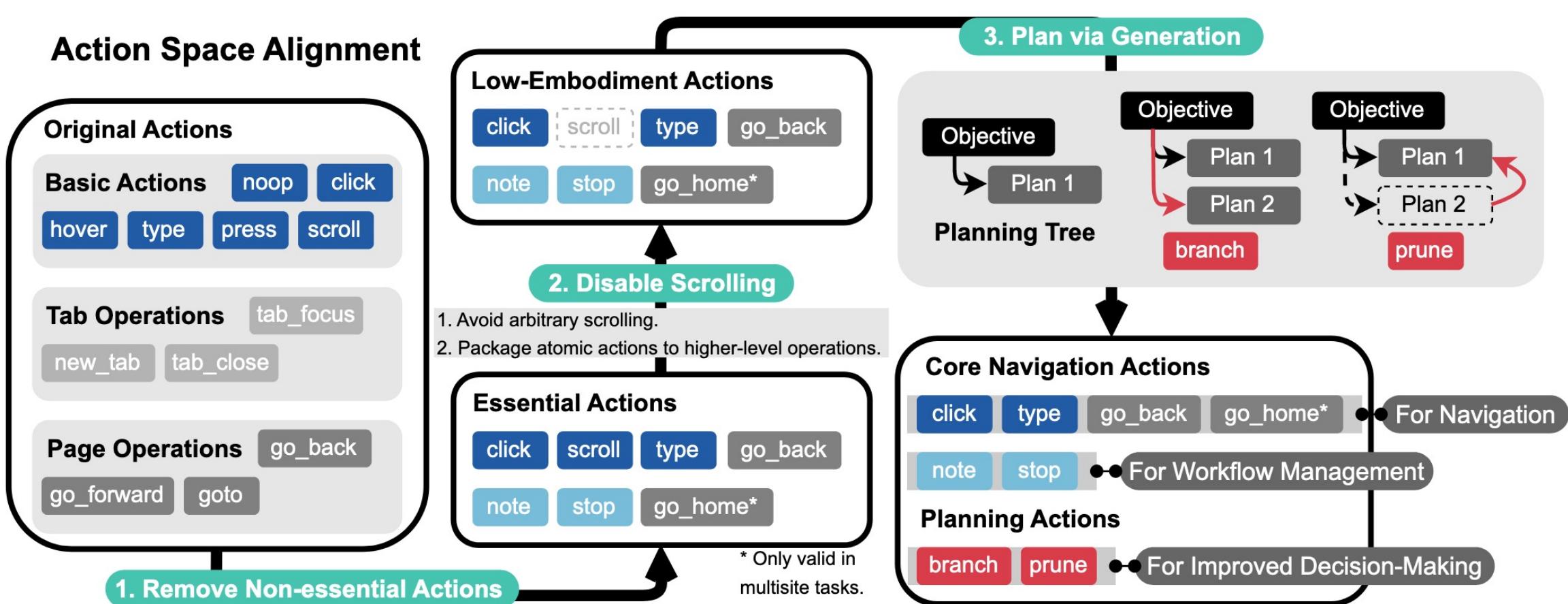


Tools

Interacting – Web Browsing



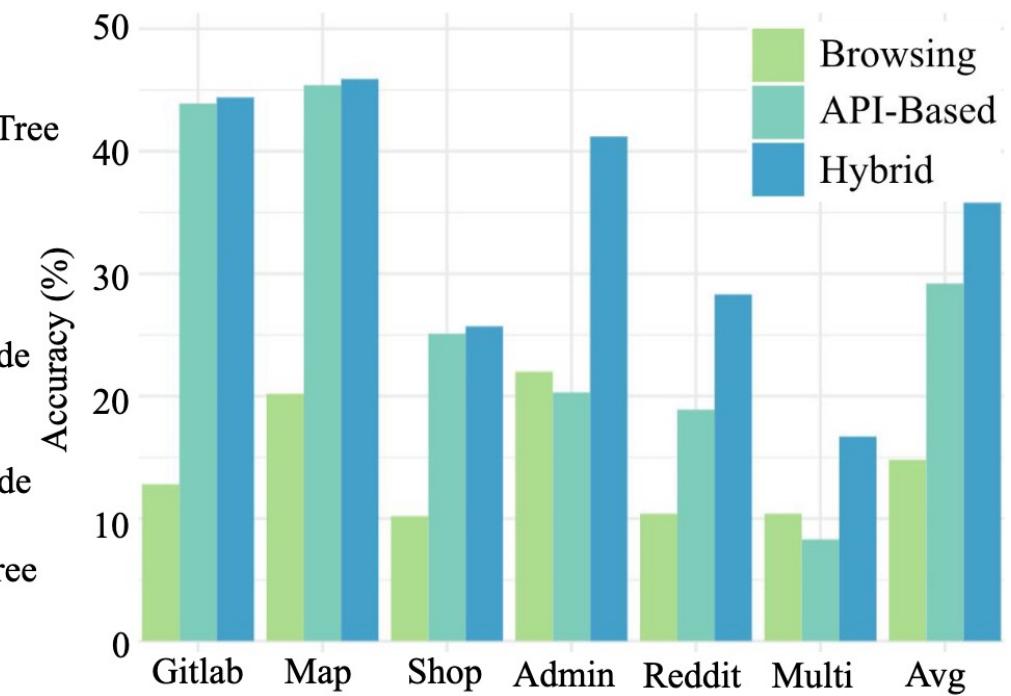
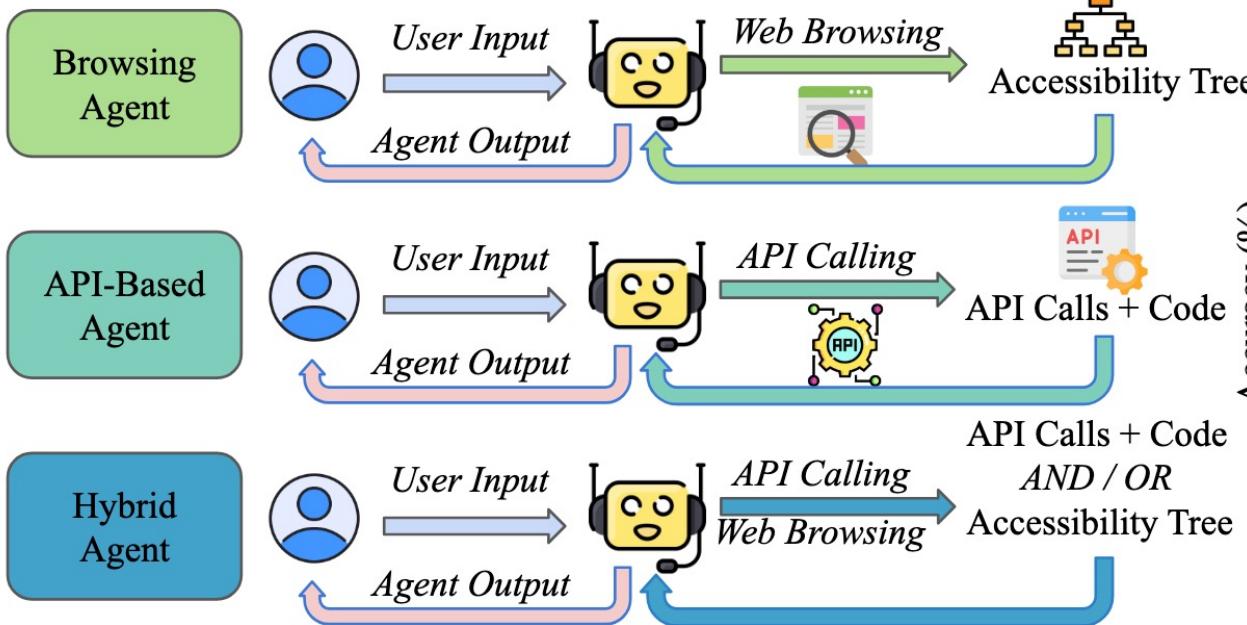
☐ AGENTOCCAM



Interacting – Tool

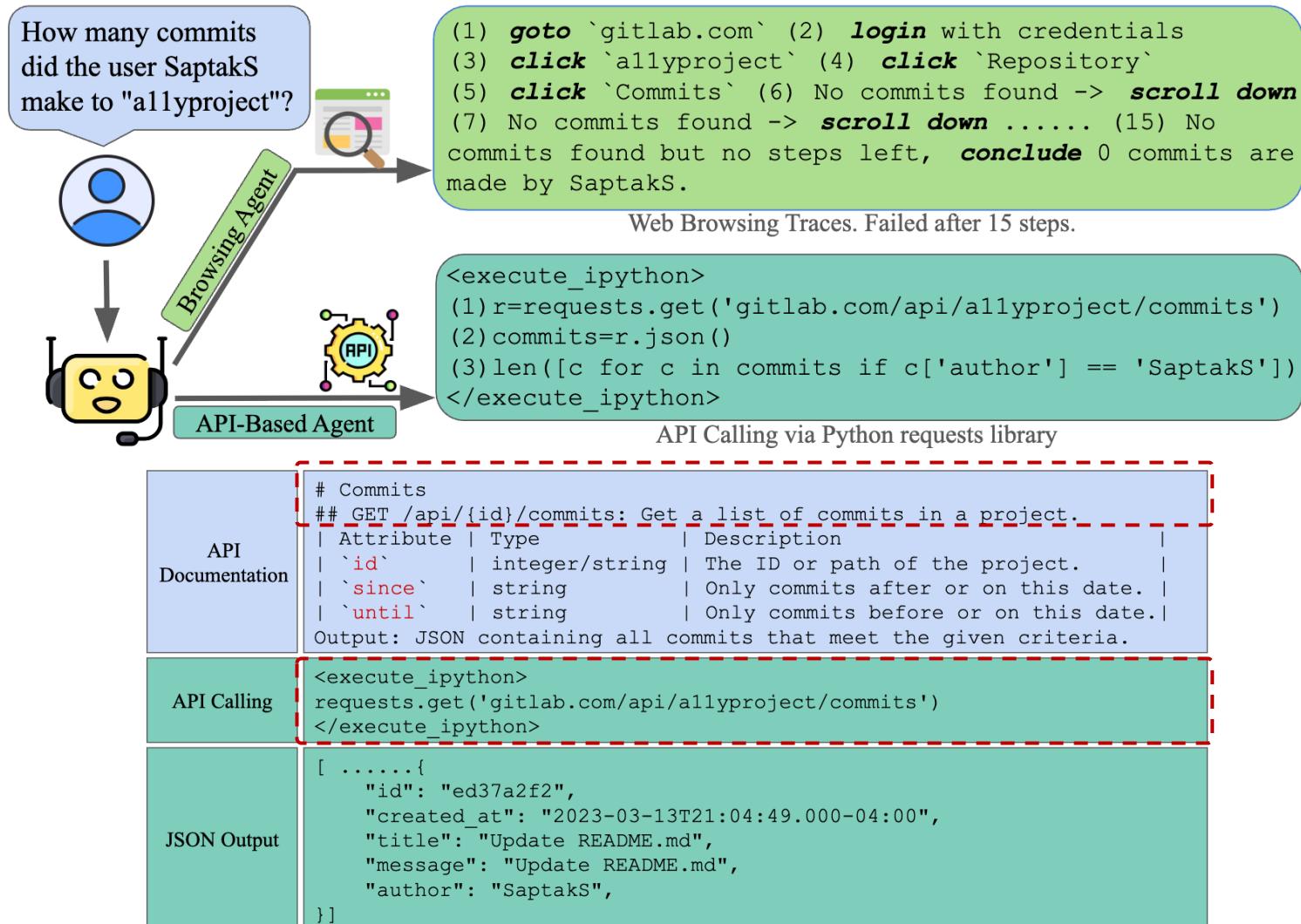
□ API-Calling Agents

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



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
- LFM need to learn how to utilize the additional tool
- ...