THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學 | MICHIGAN STATE UNIVERSITY | UIC UNIVERSITY OF ILLINOIS CHICAGO | KDD2025 August 3–7, 2025

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


Yujuan Ding[1]


Liangbo Ning[1]


Ziran Liang[1]


Zhuohang Jiang[1]


Haohao Qu[1]


Wenqi Fan[1]


Qing Li[1]


Hui Liu[2]


Xiaoyong Wei[1]


Philip S. Yu[3]

[1]The Hong Kong Polytechnic University [2]Michigan State University
[3]University of Illinois at Chicago

WebSite

Survey

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

Website (Slides): https://biglemon-ning.github.io/WebAgents/
Survey Paper: https://arxiv.org/abs/2503.23350

Tutorial Outline

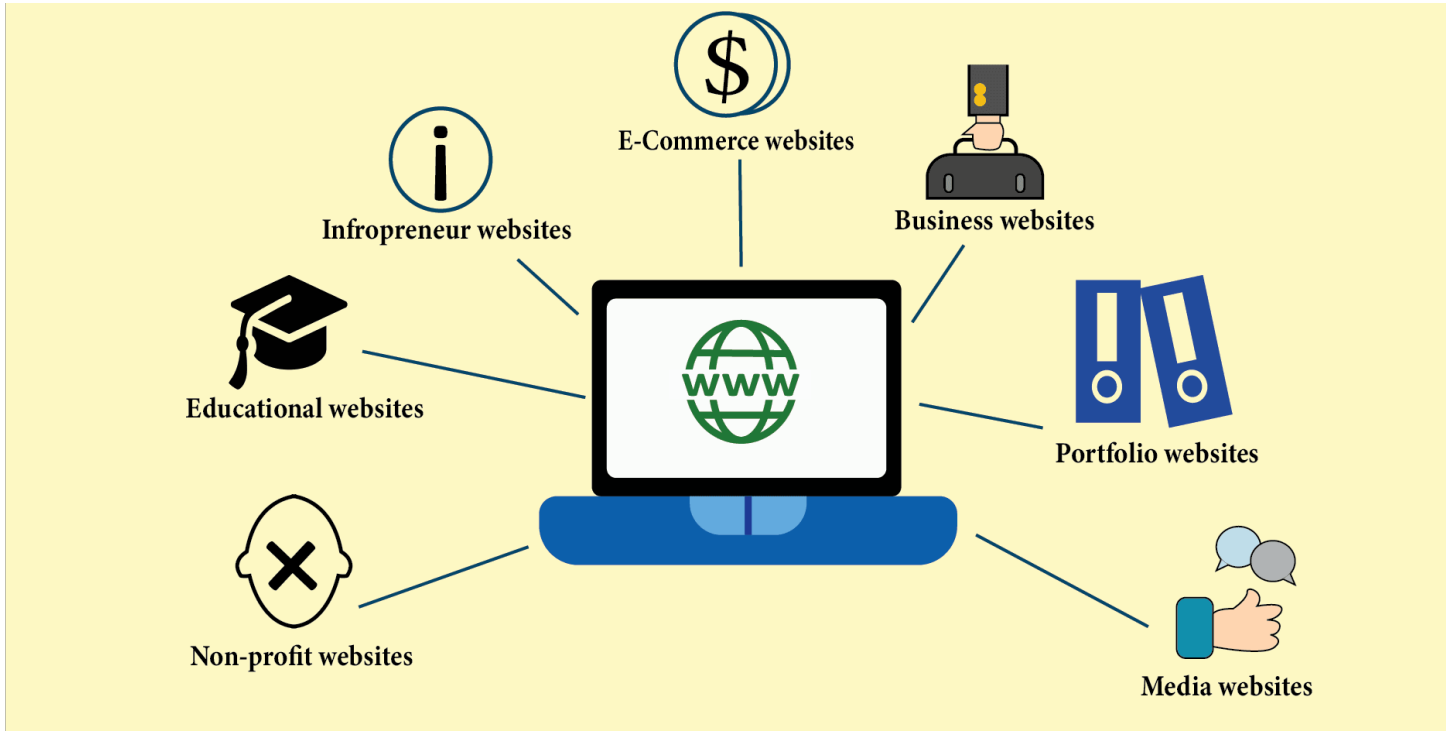
- **Part 1: Introduction of WebAgents (Yujuan Ding)**
- **Part 2: Preliminaries** of AI Agents and LFM-based WebAgents (Zhuohang Jiang)
- **Part 3: Architectures** of WebAgents (Yujuan Ding)
- **Coffee Break**
- **Part 4: Training of WebAgents** (Yujuan Ding)
- **Part 5: Trustworthy WebAgents** (Haohao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

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

Web

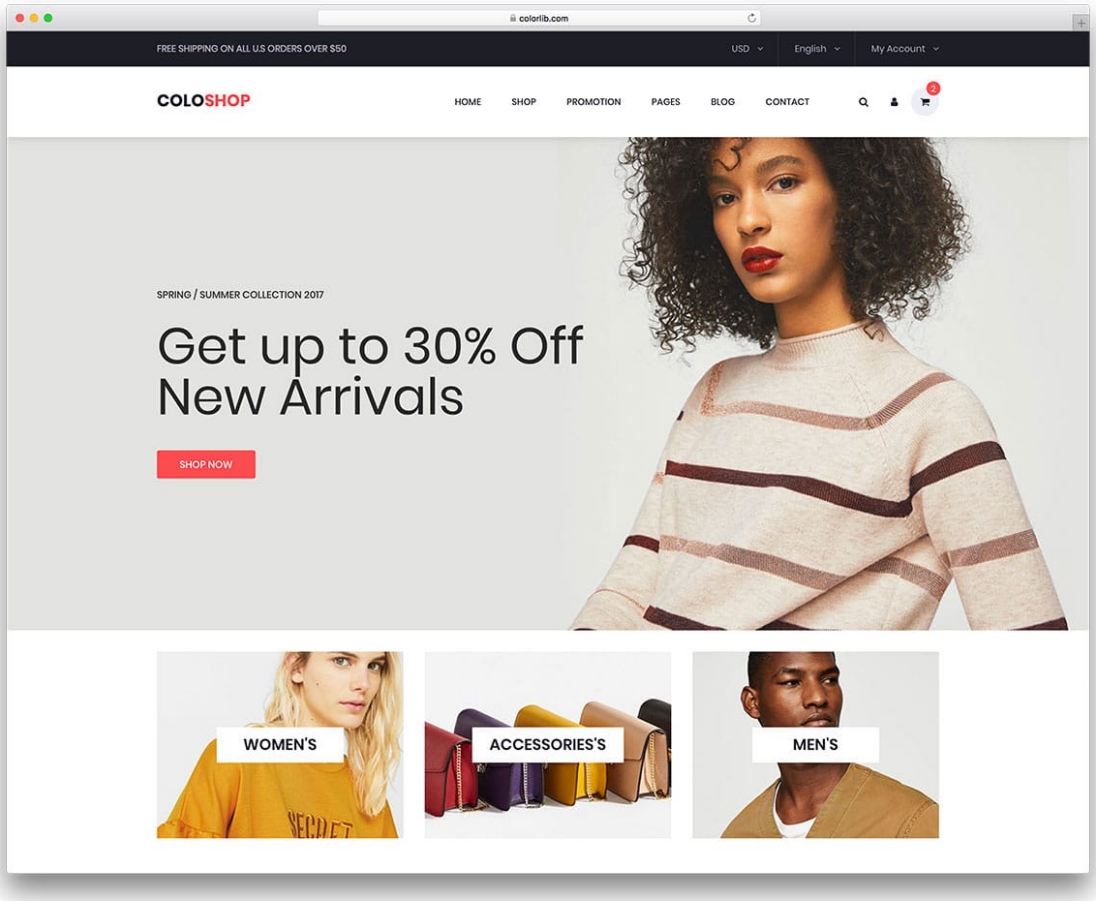
"The Web does not just connect machines, it connects people."

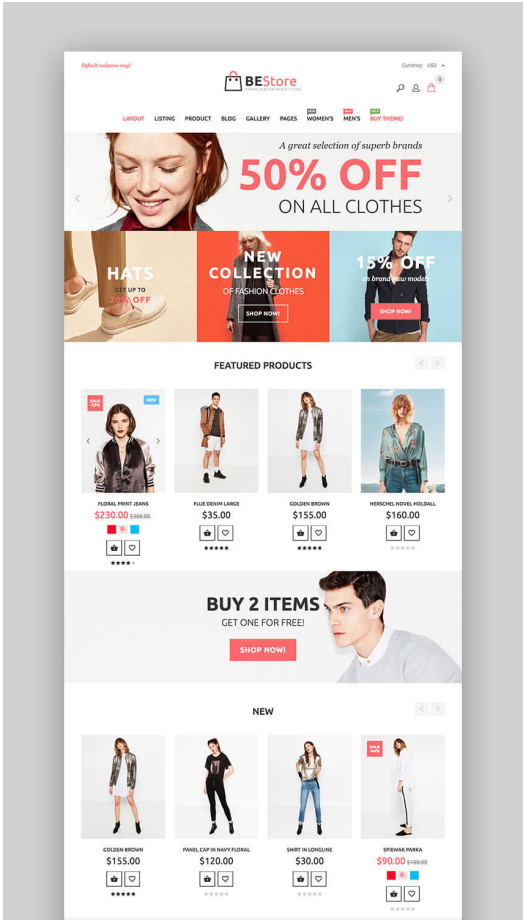
—Tim Berners-Lee,
Inventor of the World Wide Web,
ACM A.M. Turing Award Laureate 2016.



Web for E-Commence

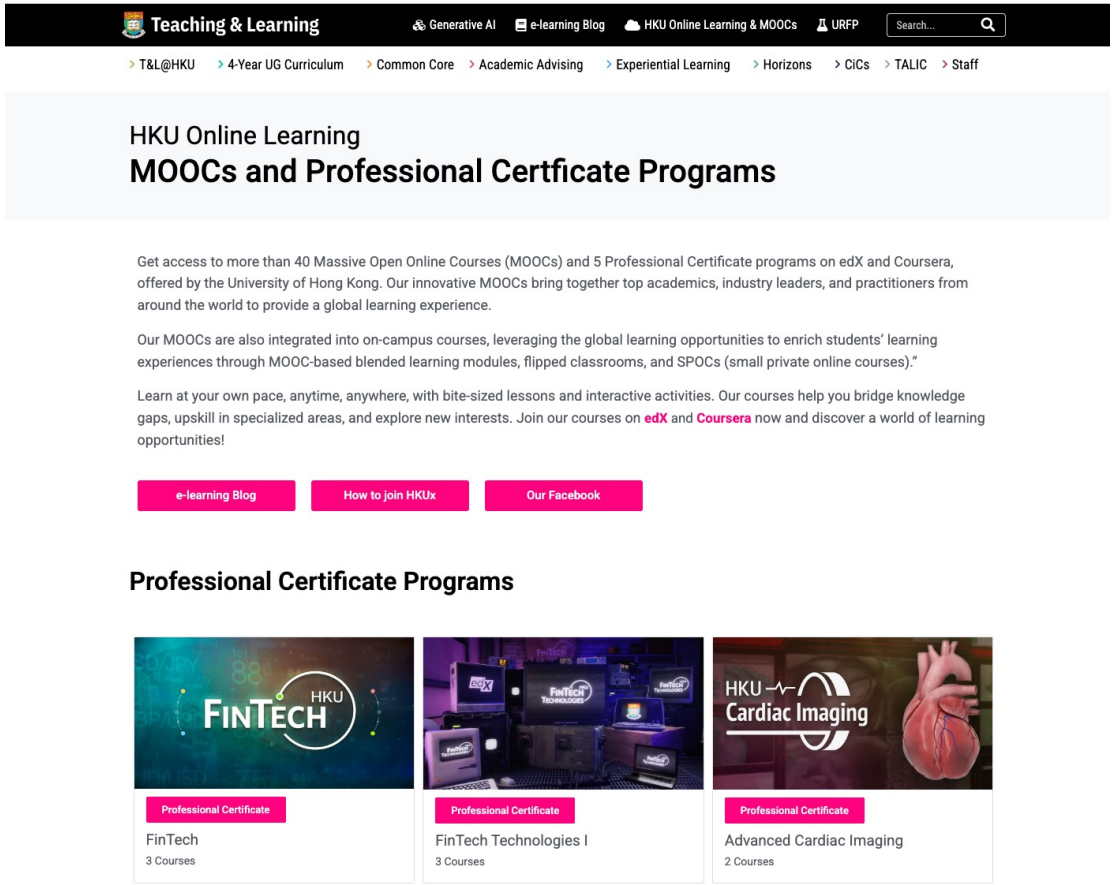
Web provides a digital platform that enables businesses and consumers to buy and sell goods and services online through websites and applications.





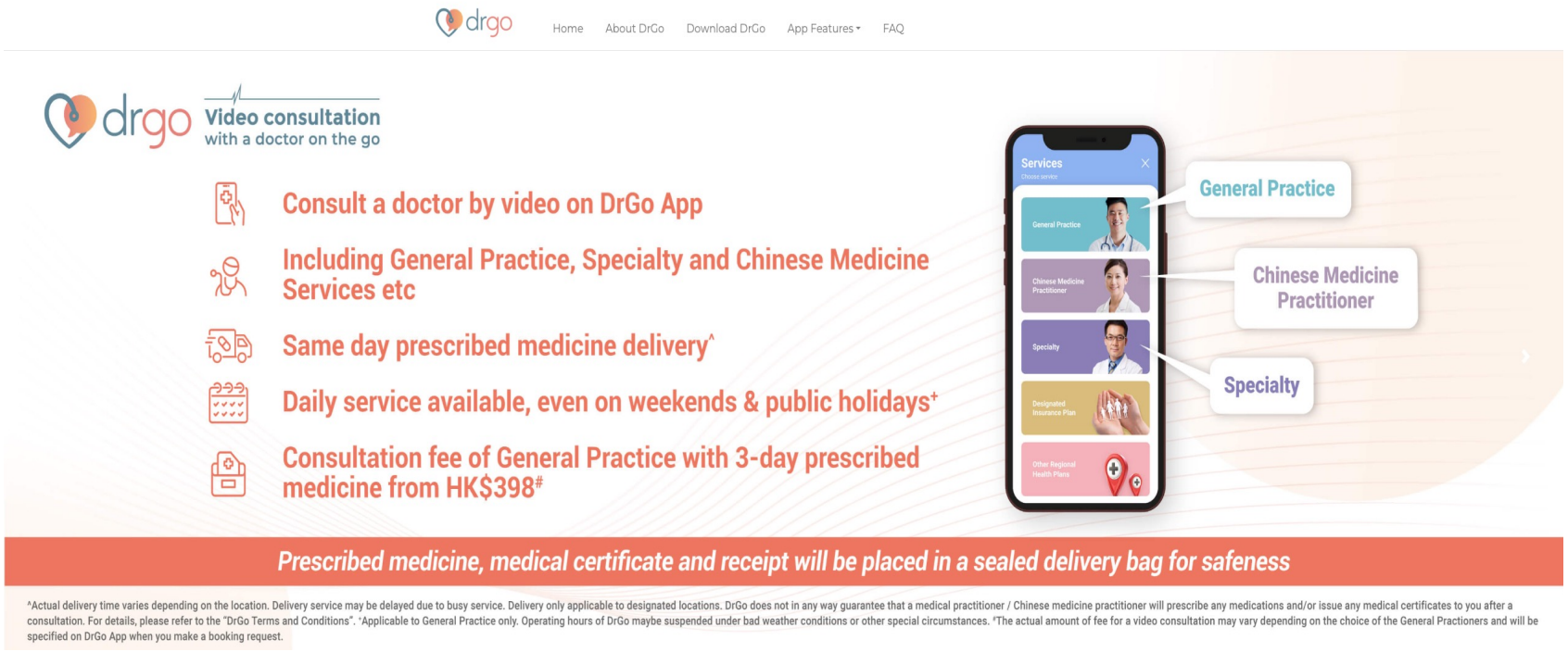
Web for Education

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



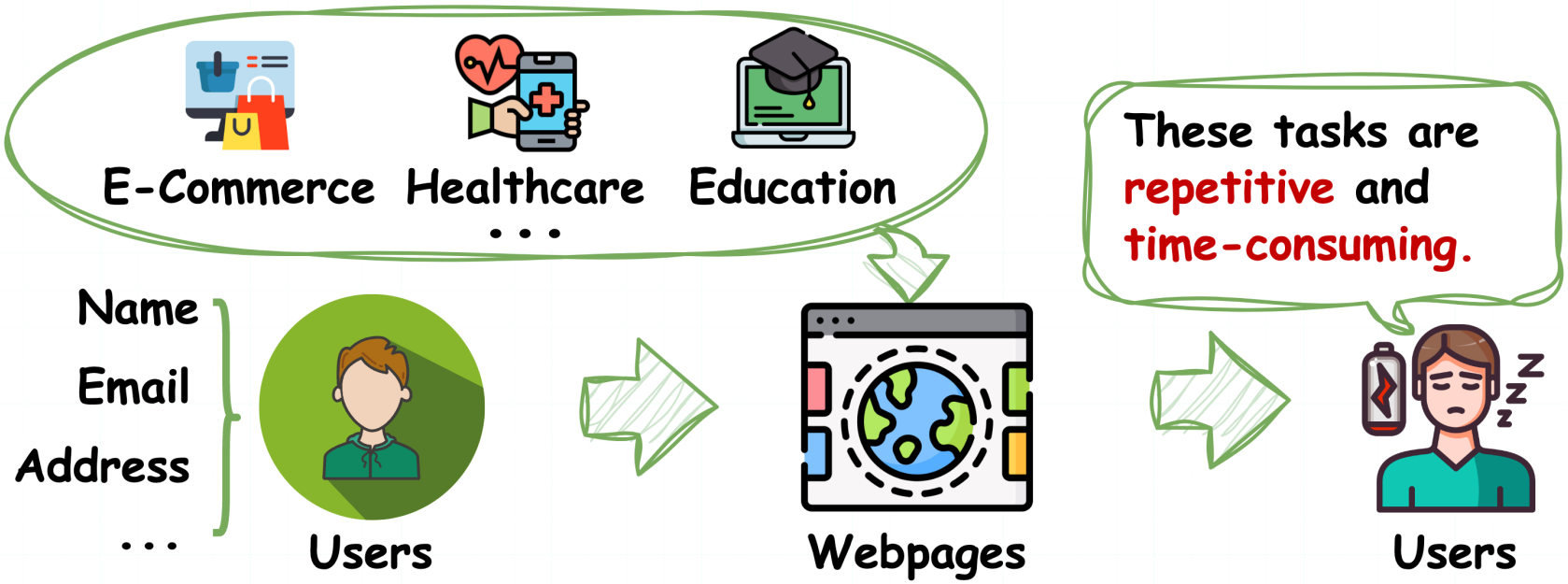
Web for Healthcare

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



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.



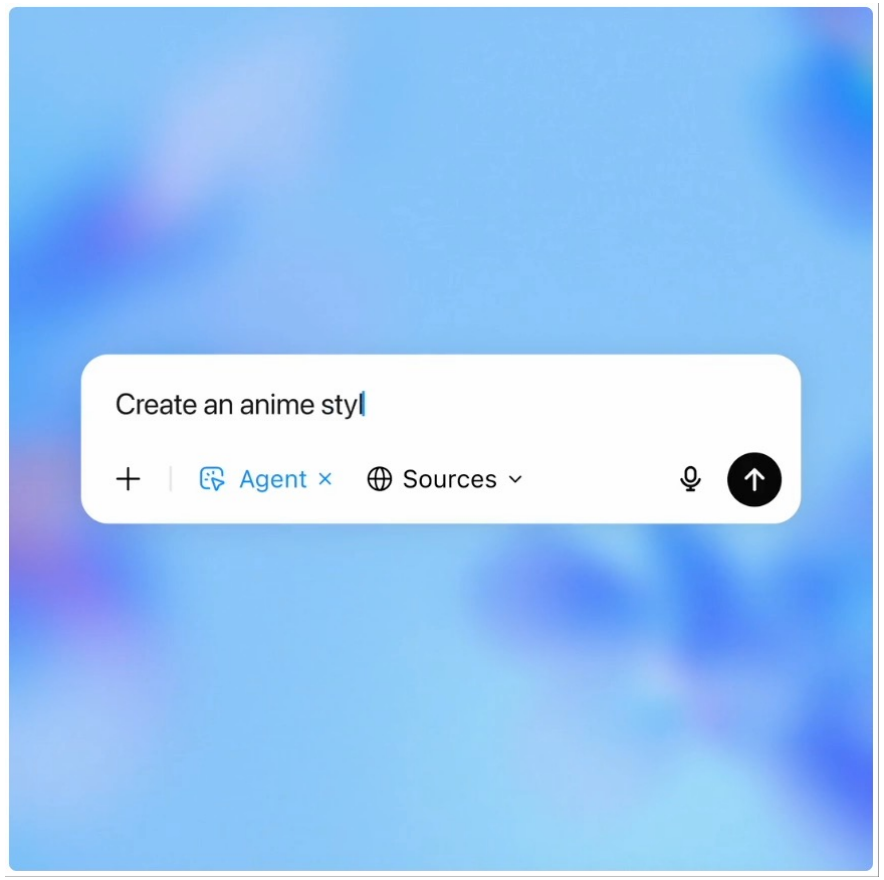
AI Agents for Web

"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

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

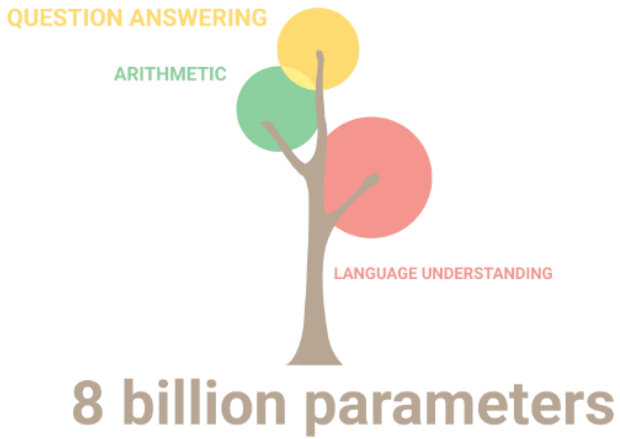
1,997,406 plays | Sam Altman | TED2025 • April 2025



Large Foundations Models (LFMs)

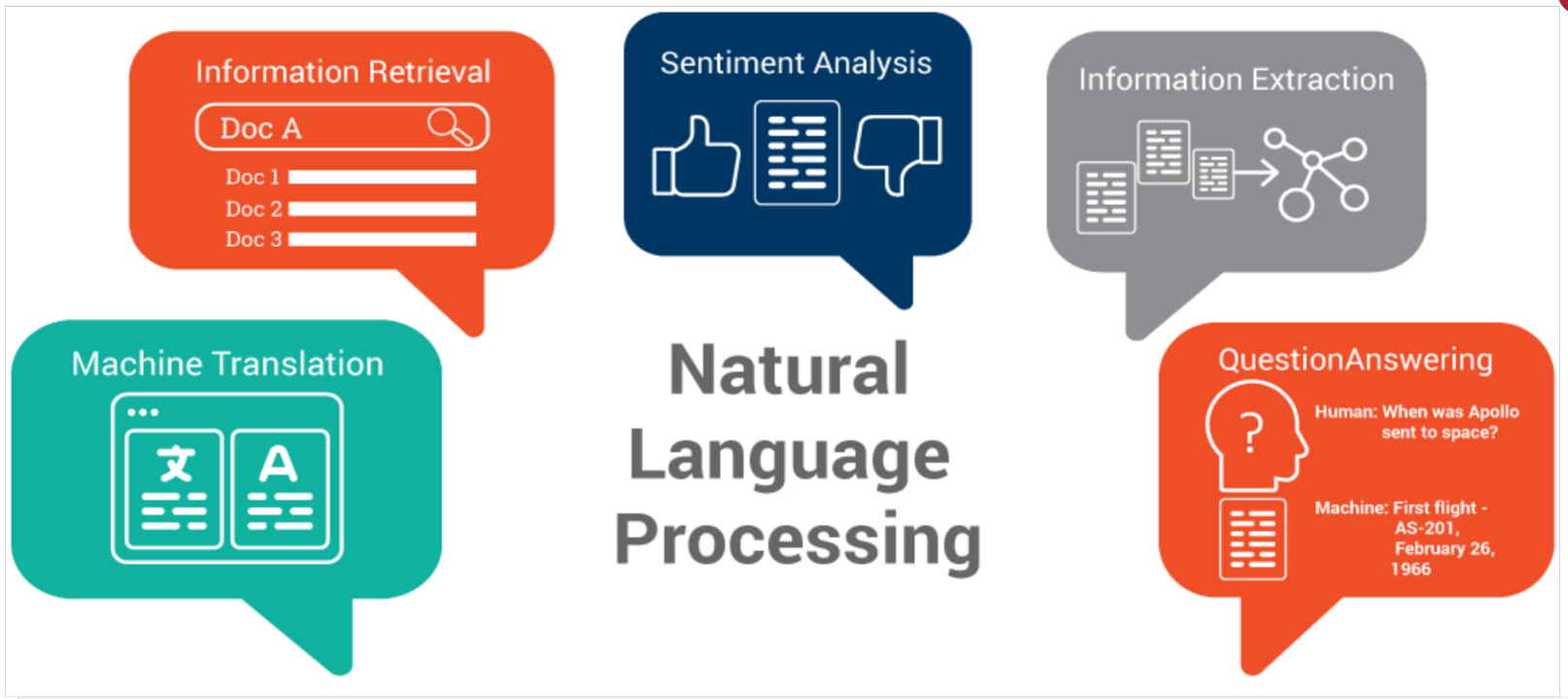
They Are Changing Our Lives !

OpenAI | deepseek | Meta | Claude | Qwen .....



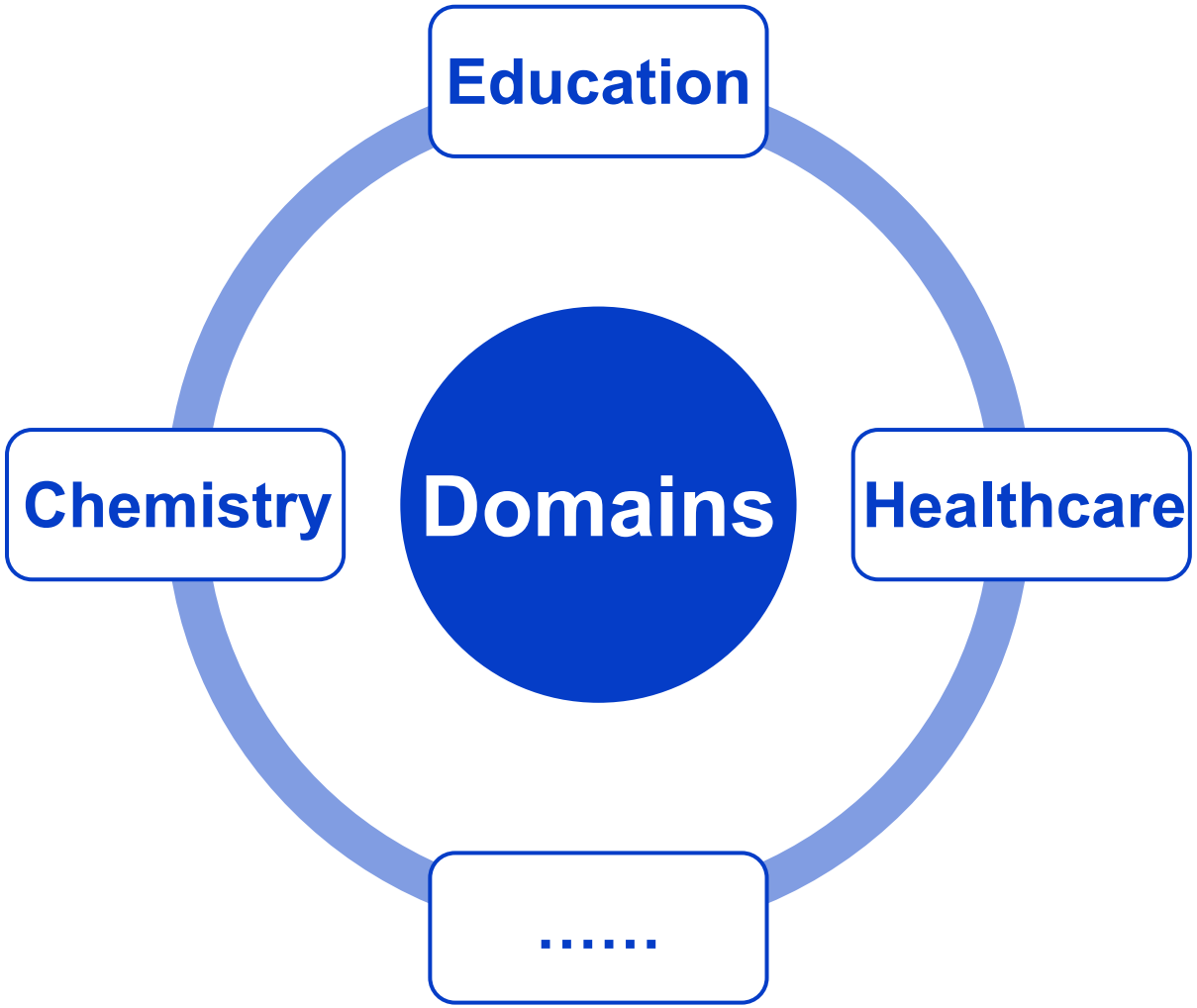
https://github.com/Hannibal046/Awesome-LLM/tree/main

LFMs in Natural Language Processing

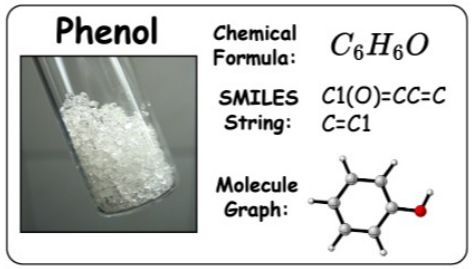


Large Language Models (LLMs)

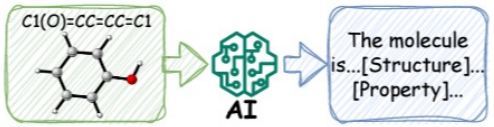
LFMs in Downstream Domains



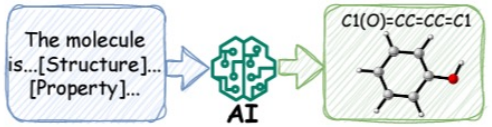
❑ **Molecule discovery**, etc.

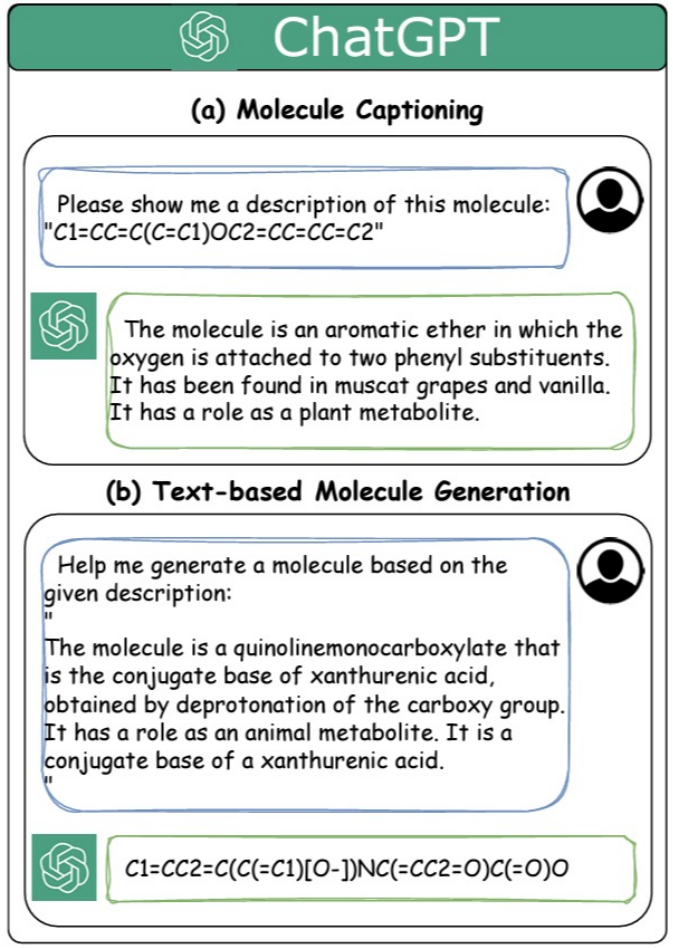


(a) Molecule Representations.



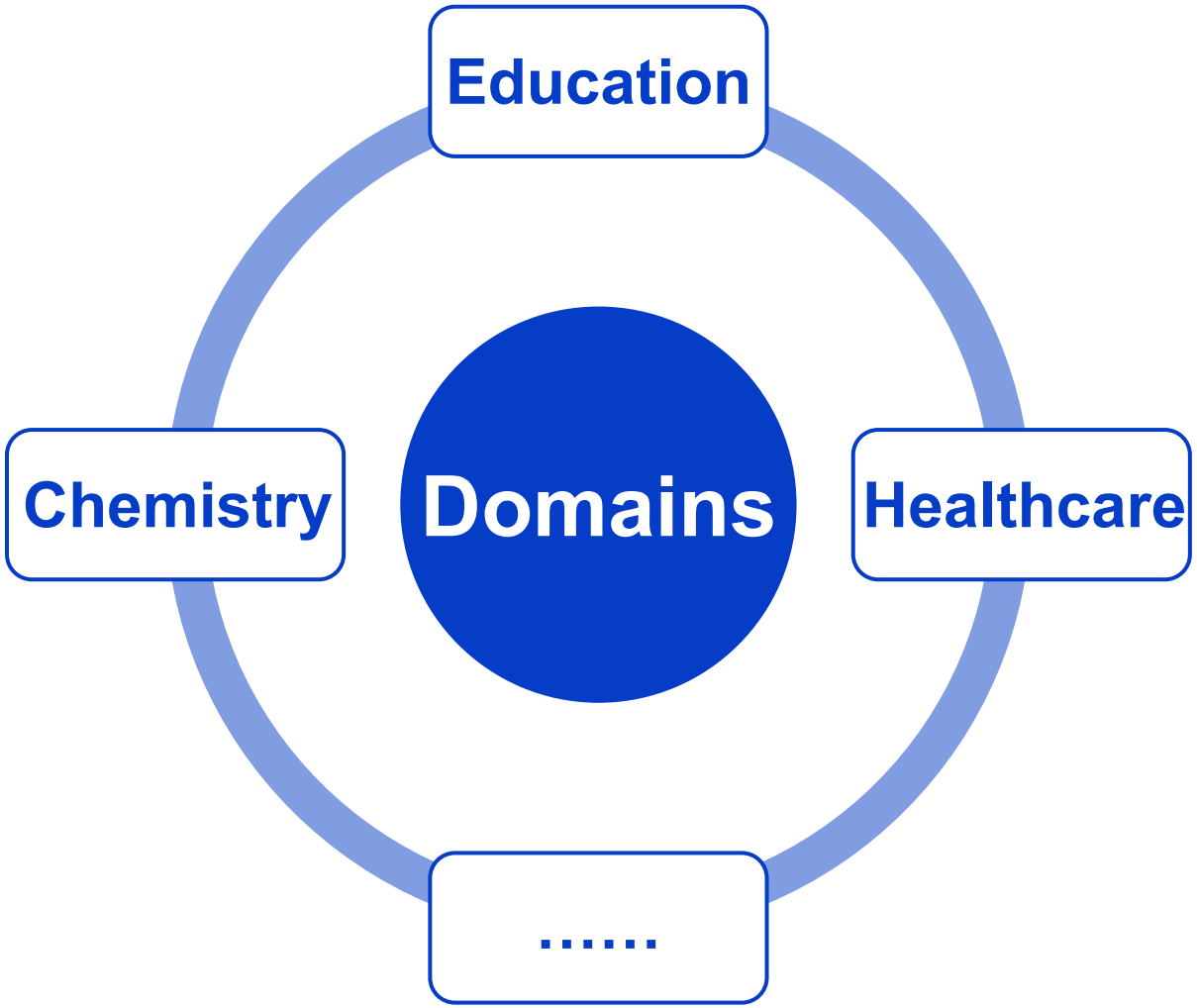
(b) Molecule Captioning.



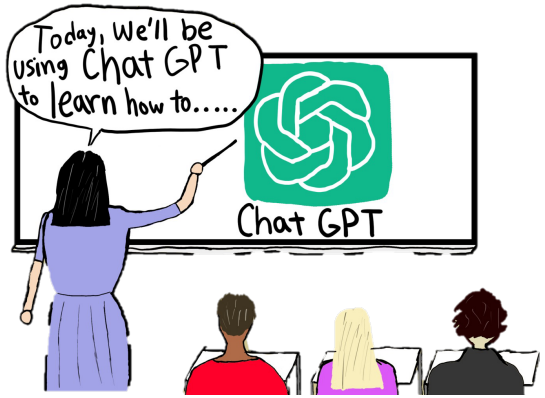


Li, Jiatong, et al. "Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective." *TKDE,* 36.11 (2024): 6071-6083.
Liu, Yuyan, et al. "Moleculargpt: Open large language model (llm) for few-shot molecular property prediction." *arXiv preprint arXiv:2406.12950* (2024).

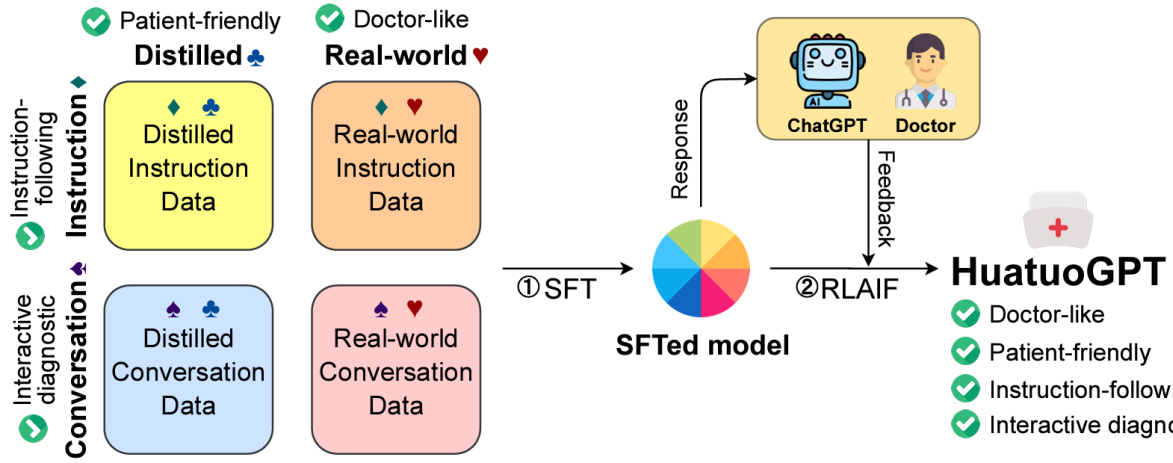
LFMs in Downstream Domains



❑ **Curriculum & Teaching**, etc.



❑ **Medical consultation**, etc.



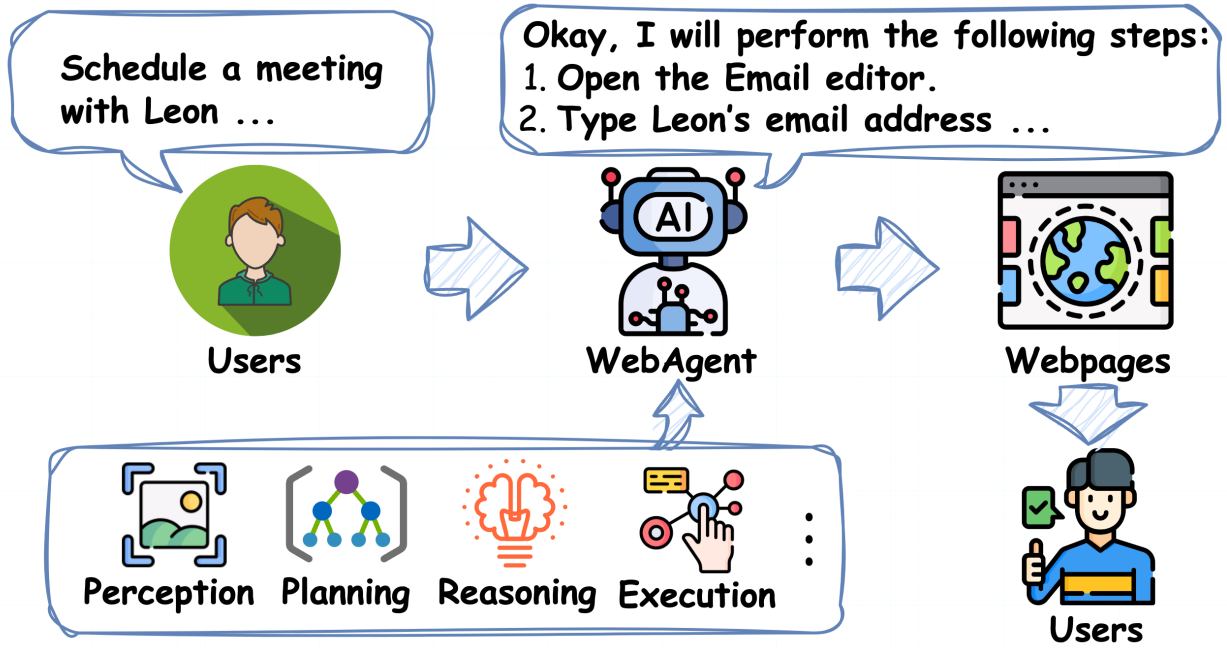
Zhang, Hongbo, et al. "HuatuoGPT, Towards Taming Language Model to Be a Doctor." EMNLP, 2023.

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.*

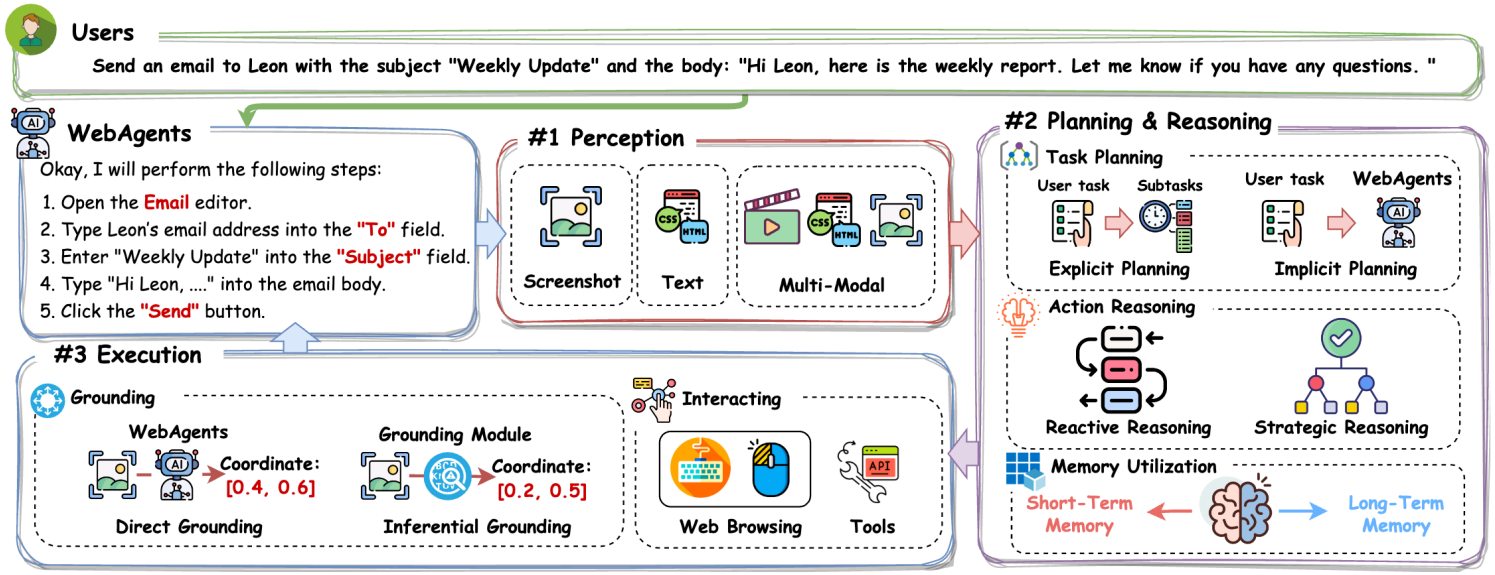


A Comprehensive Survey Paper

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

Liangbo Ning[1], Ziran Liang[1], Zhuohang Jiang[1], Haohao Qu[1], Yujuan Ding[1], Wenqi Fan[1]*, Xiao-yong Wei[1], Shanru Lin[2], Hui Liu[3], Philip S. Yu[4], Qing Li[1]*
[1]The Hong Kong Polytechnic University, [2]City University of Hong Kong,
[3]Michigan State University, [4]University of Illinois at Chicago

https://arxiv.org/pdf/2503.23350



Survey paper on KDD

Tutorial Website (Slides)

Survey

WebSite

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://wenqifan03.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

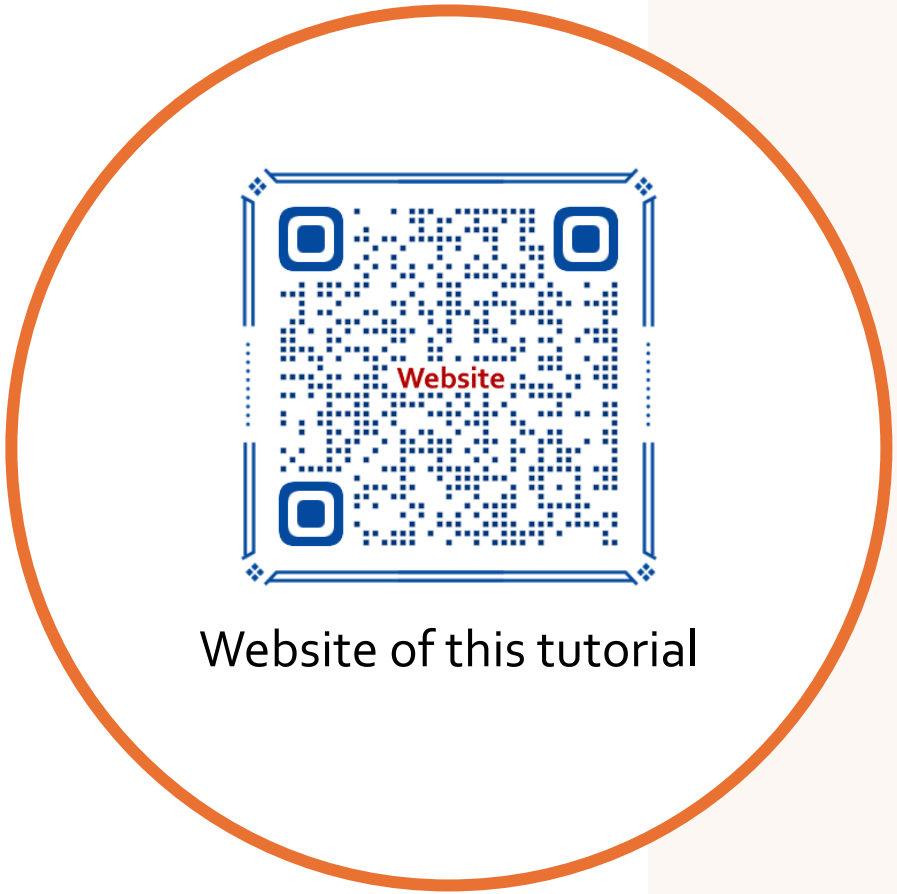
- AI Agents with human-like reasoning and autonomous decision-making capabilities have revolutionized various domains.
  - Reinforcement learning (RL)-based Agents
  - LFM-empowered Agents.

Medicine

Finance

Education

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

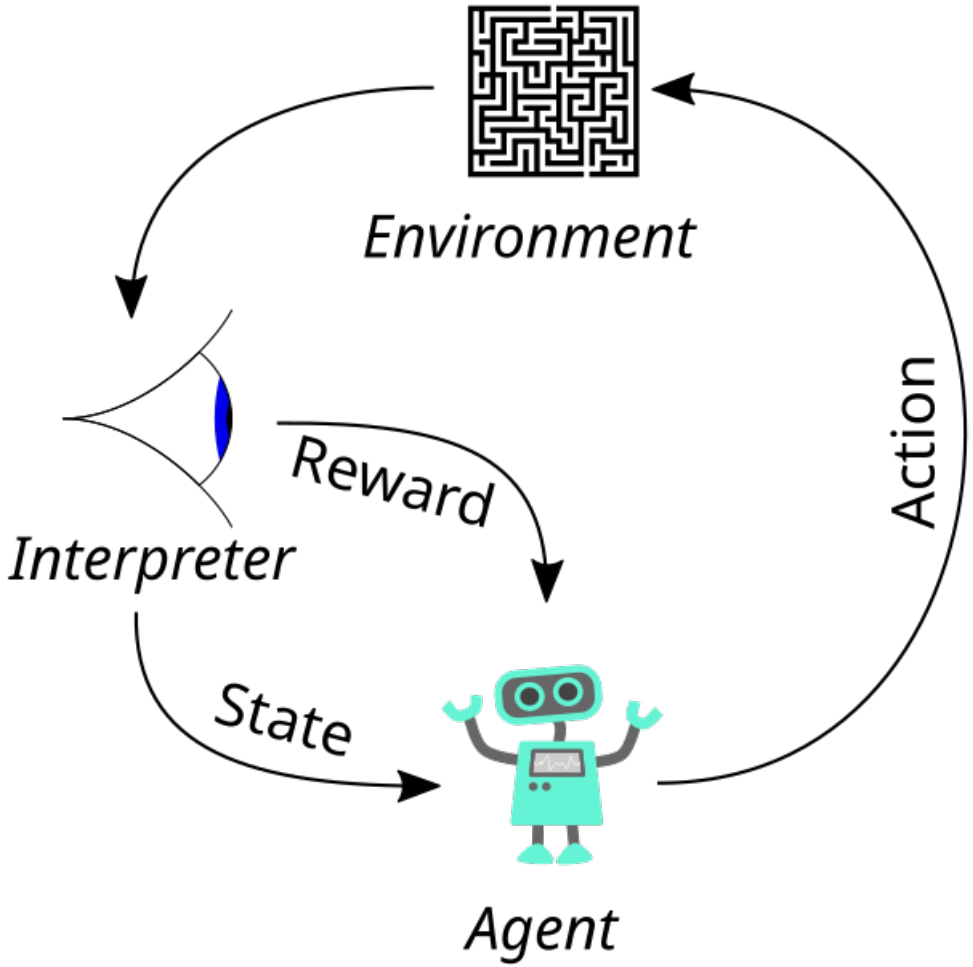


Website of this tutorial

- 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

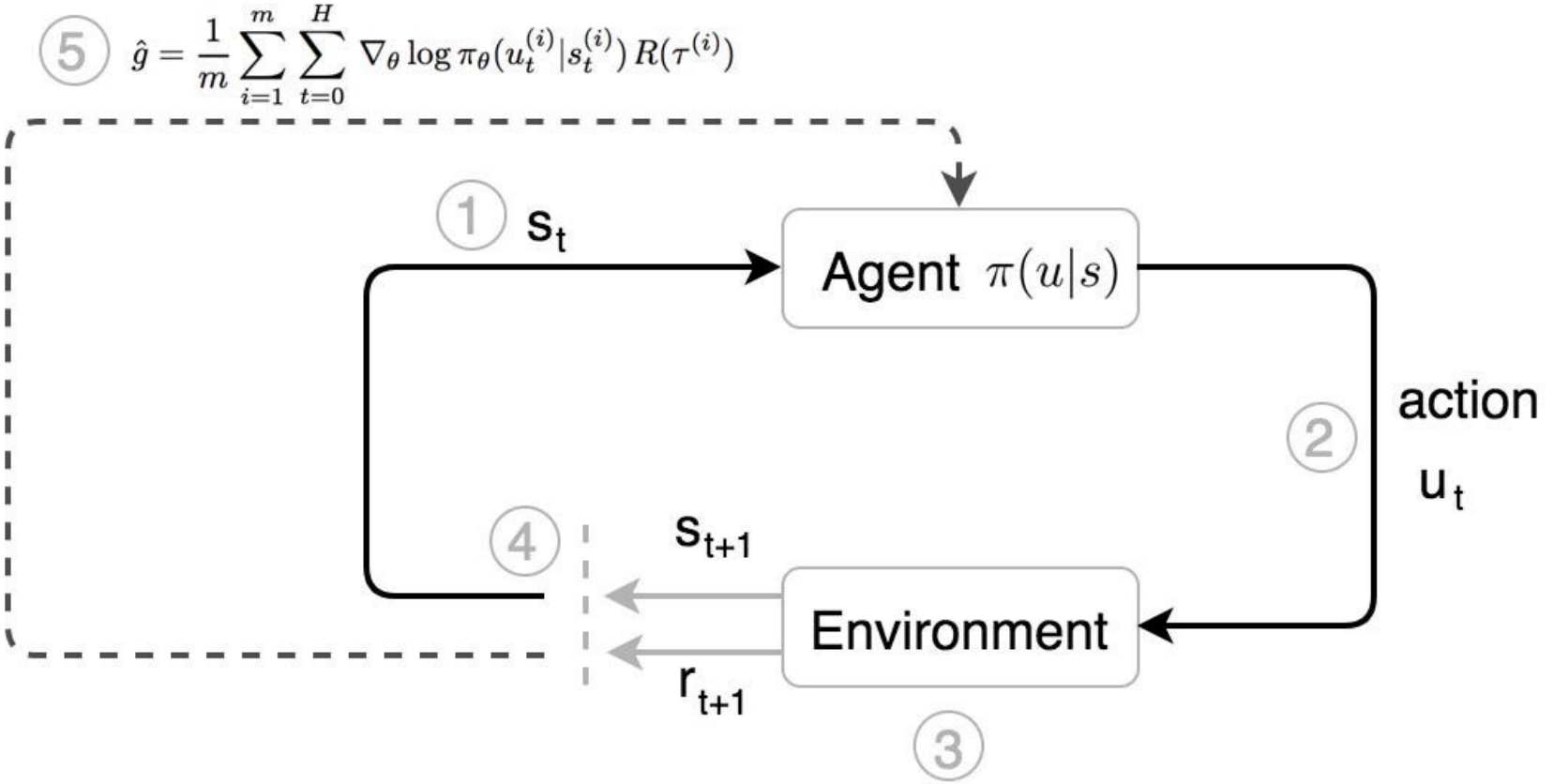
- **Q-Learning** learns the optimal action-value function by iteratively updating Q-values based on rewards and future estimates.



Watkins, Christopher JCH, and Peter Dayan. "Q-learning." Machine learning 8.3 (1992): 279-292.

Reinforcement Learning

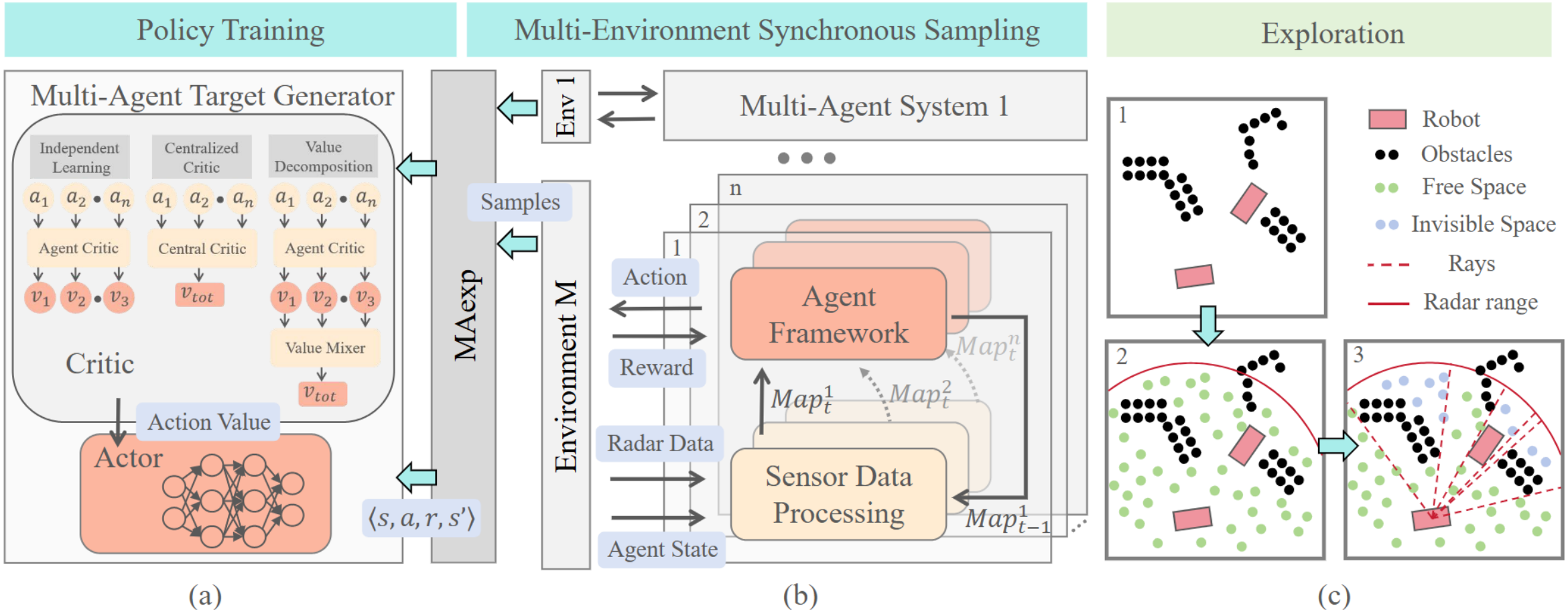
- **Policy Gradient** directly optimizes actions by maximizing expected rewards.



Sutton, Richard S., et al. "Policy gradient methods for reinforcement learning with function approximation." Advances in neural information processing systems 12 (1999).

RL-based Agents

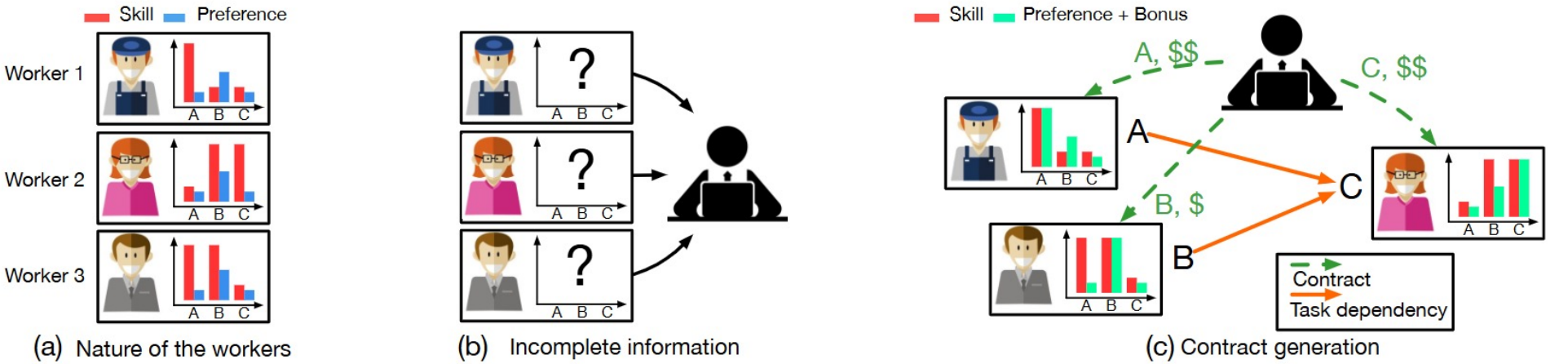
❑ Exploration (Maexp)

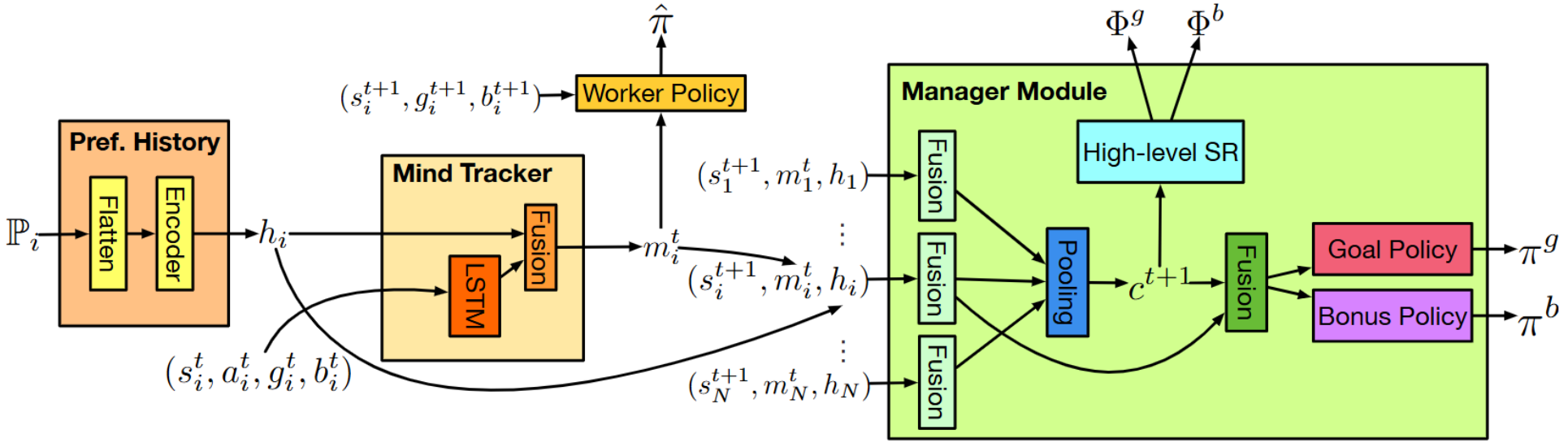


Zhu, Shaohao, et al. "Maexp: A generic platform for rl-based multi-agent exploration." 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024.

RL-based Agents

❑ Management (M3RL)



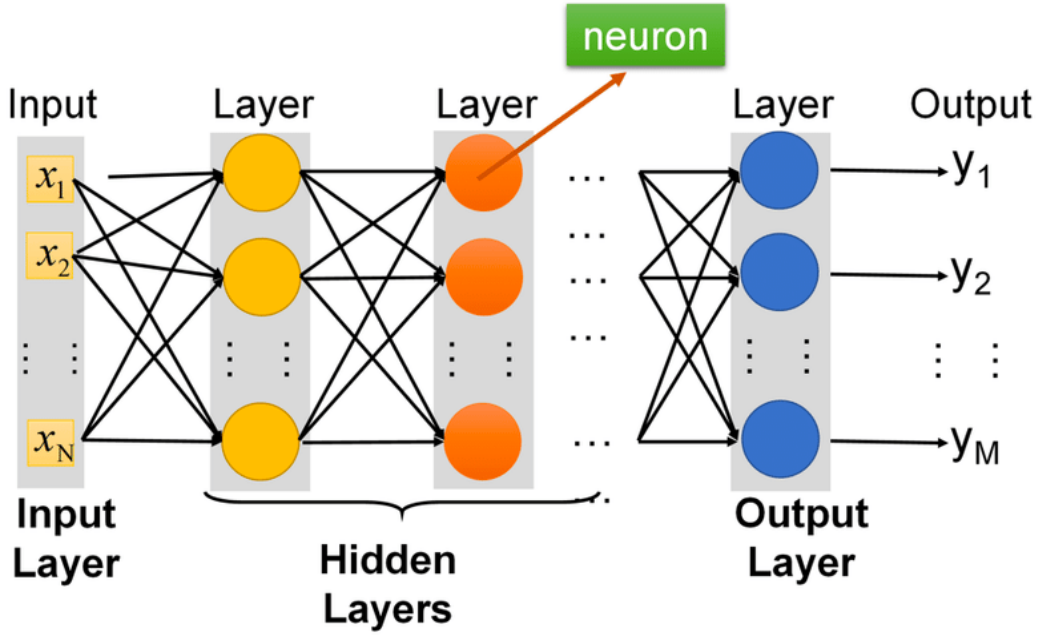


Shu, Tianmin, and Yuandong Tian. "M^ 3RL: Mind-aware Multi-agent Management Reinforcement Learning." International Conference on Learning Representations, 2019.

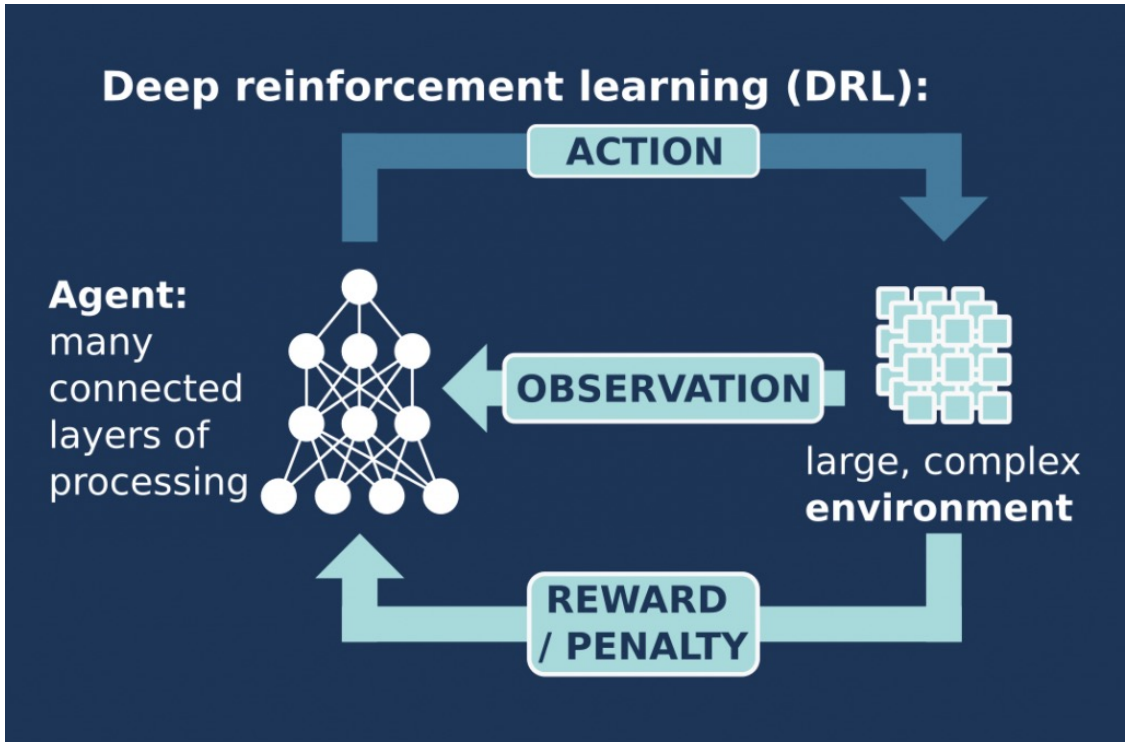
Deep Reinforcement Learning

- 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



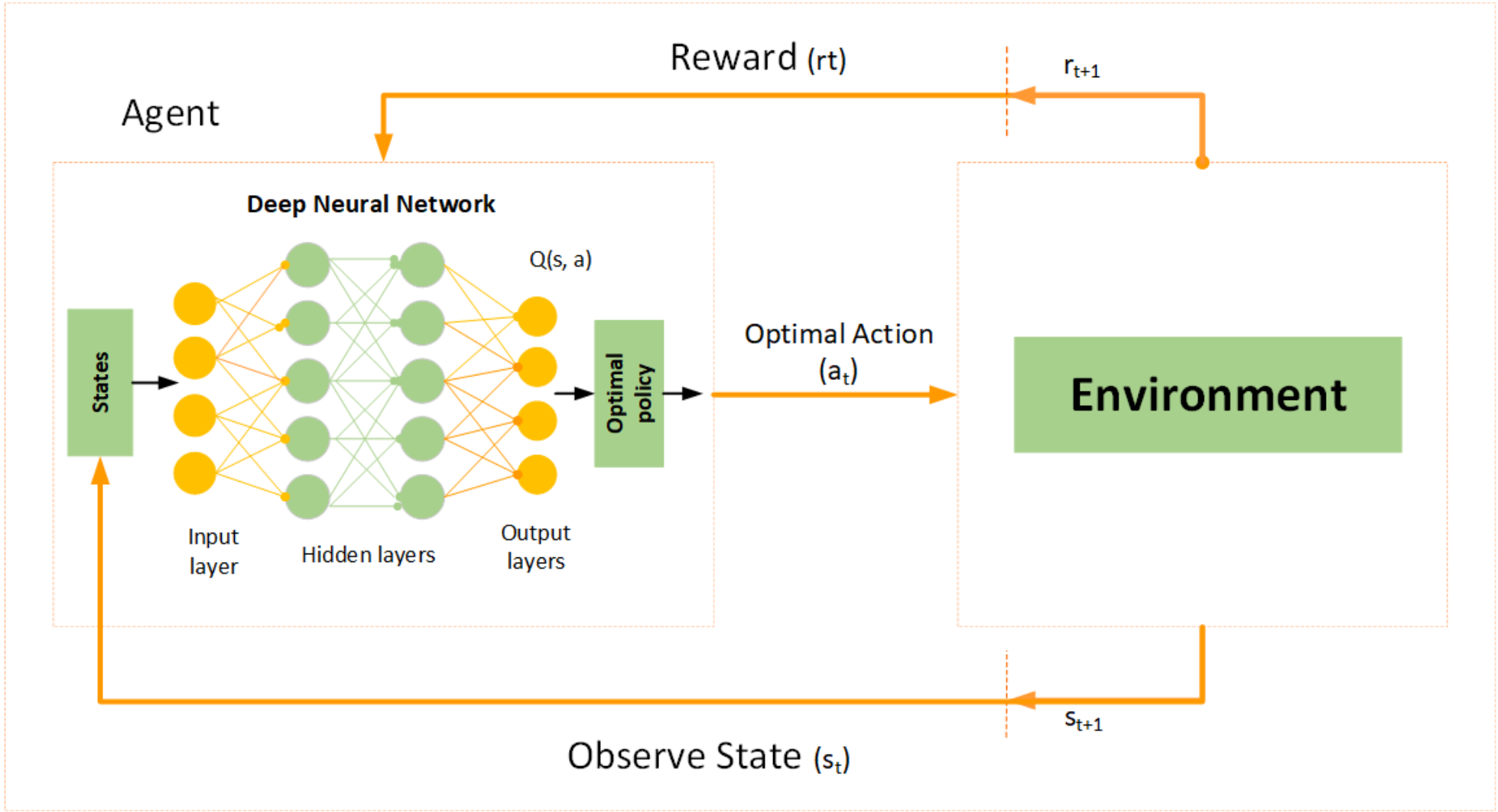
Deep Reinforcement Learning



Wang, Haozhe, et al. "Data-driven dynamic resource scheduling for network slicing: A deep reinforcement learning approach." Information Sciences 498 (2019): 106-116.

Deep Reinforcement Learning

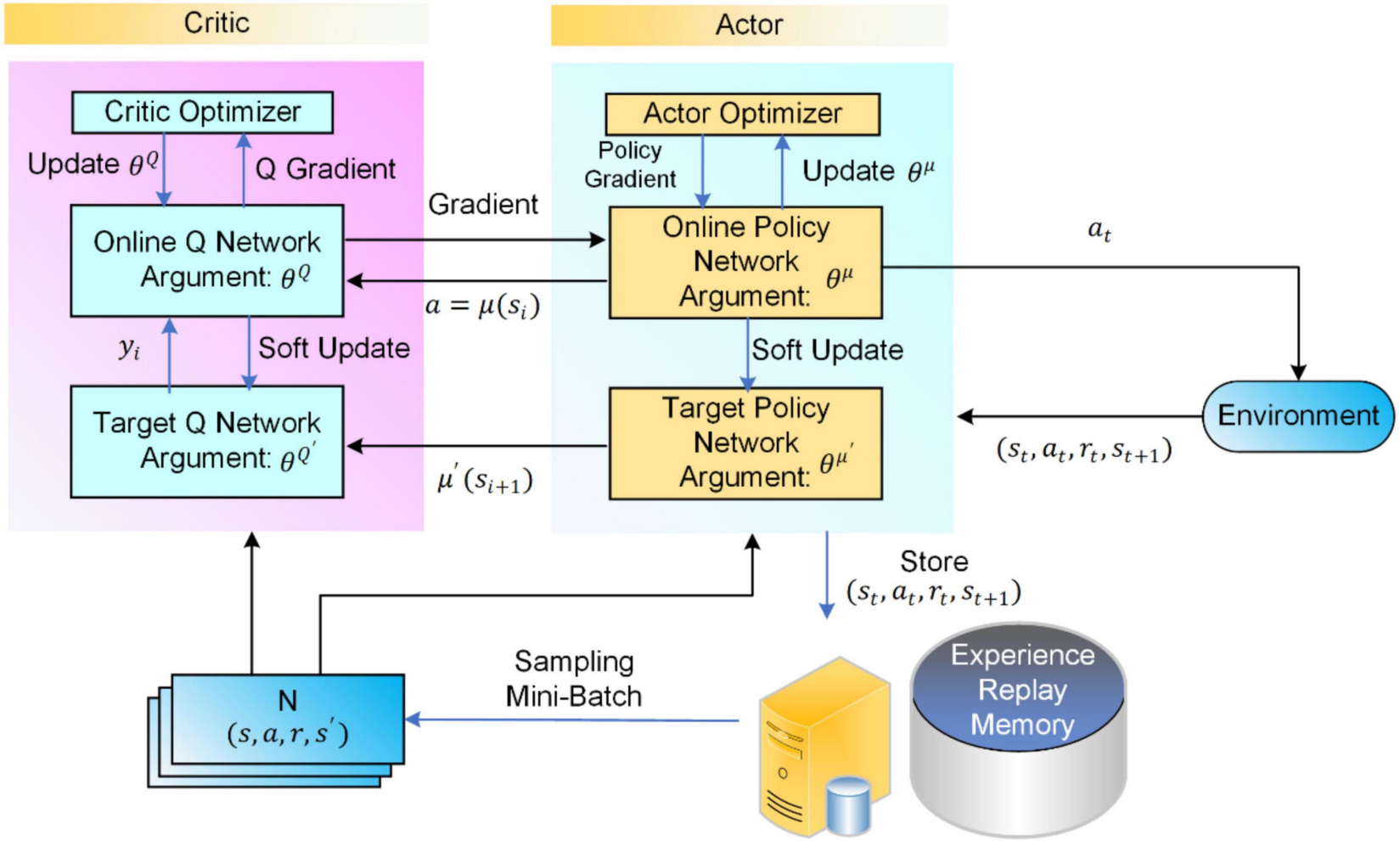
❑ Deep Q Network



Fan, Jianqing, et al. "A theoretical analysis of deep Q-learning." Learning for dynamics and control. PMLR, 2020.

Deep Reinforcement Learning

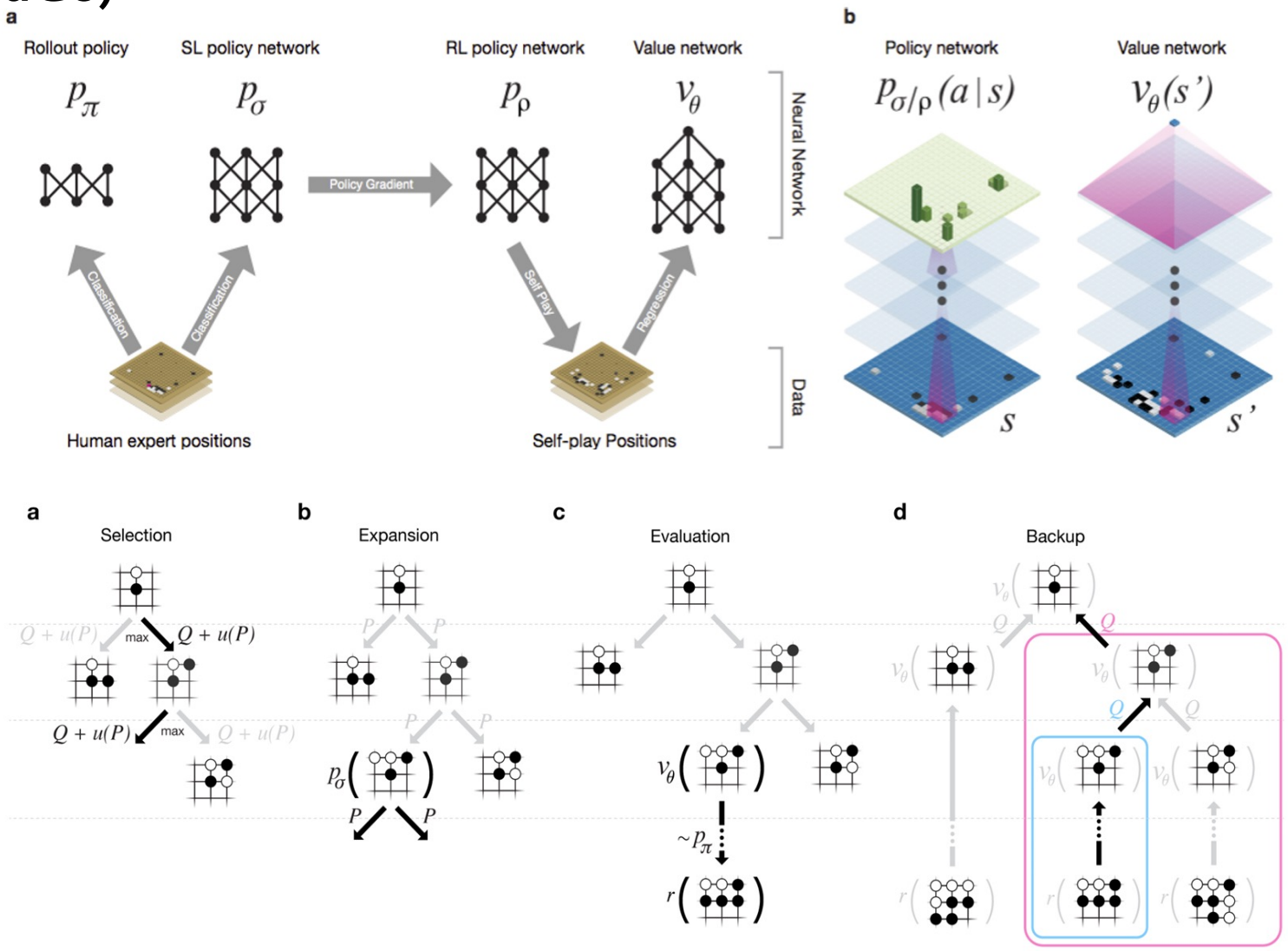
❑ Deep Deterministic Policy Gradient



Lillicrap, Timothy P., et al. "Continuous Control With Deep Reinforcement Learning." ICLR (2016).

Deep Reinforcement Learning Agents

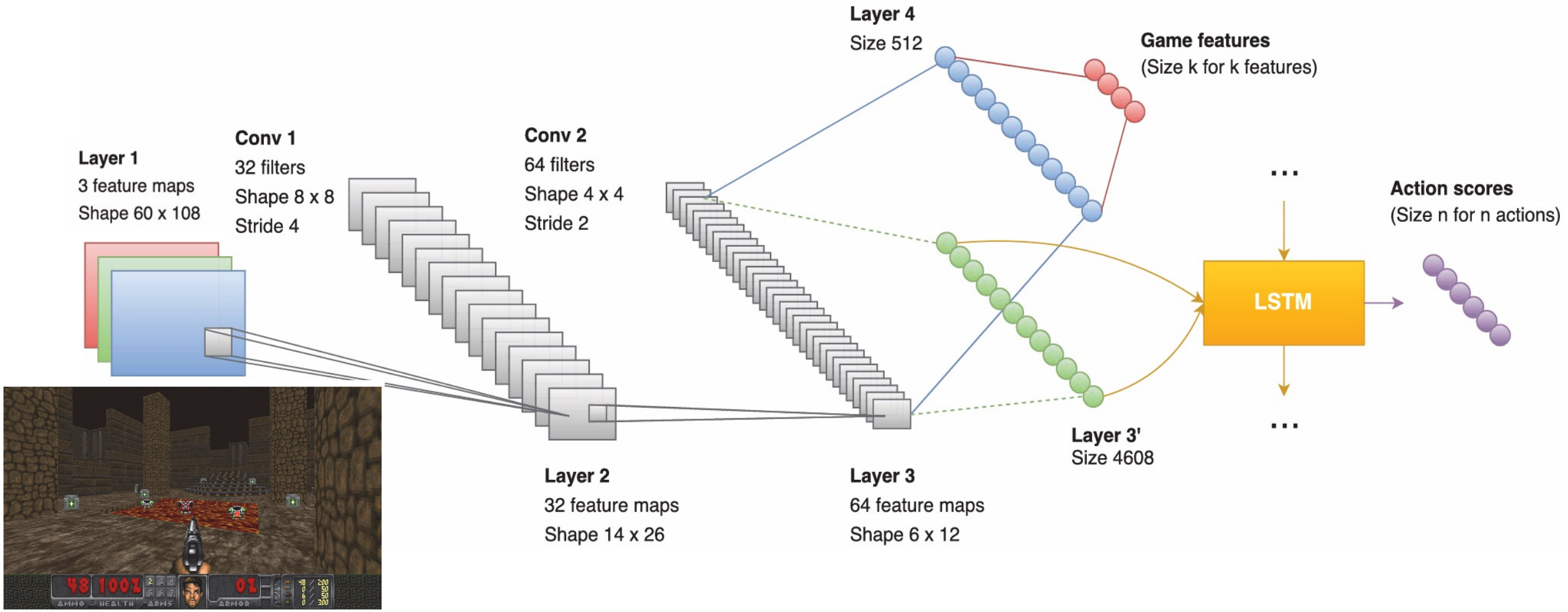
- Game (AlphaGo)



Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." nature 529.7587 (2016): 484-489.

Deep Reinforcement Learning Agents

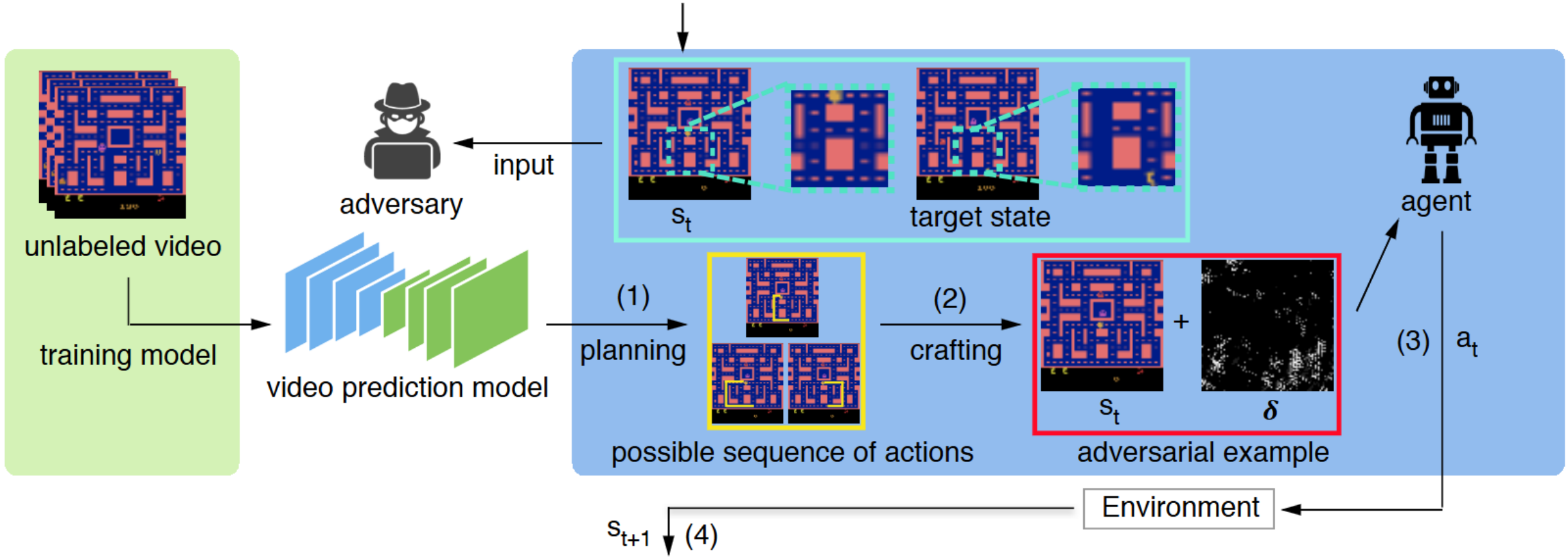
- Game (Playing FPS Games with Deep Reinforcement Learning)



Lample, Guillaume, et al. "Playing FPS games with deep reinforcement learning." Proceedings of the AAAI conference on artificial intelligence. Vol. 31. No. 1. 2017.

Deep Reinforcement Learning Agents

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



Lin, Yen-Chen, et al. "Tactics of adversarial attack on deep reinforcement learning agents." Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017.

Applications and Challenges

❑ Applications

- ➢ Game (Planning)
- ➢ Finance (Decision-making)
- ➢ …

❑ Challenges

- ➢ Lack of World Knowledge
- ➢ Lack of Adaptability
- ➢ …

Go games

Finance

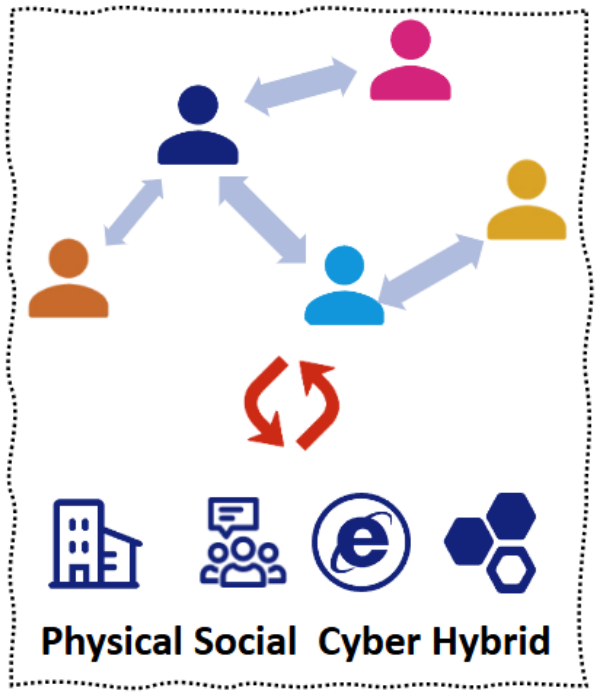
PART 2: Preliminaries of AI Agents and LFM-based WebAgents

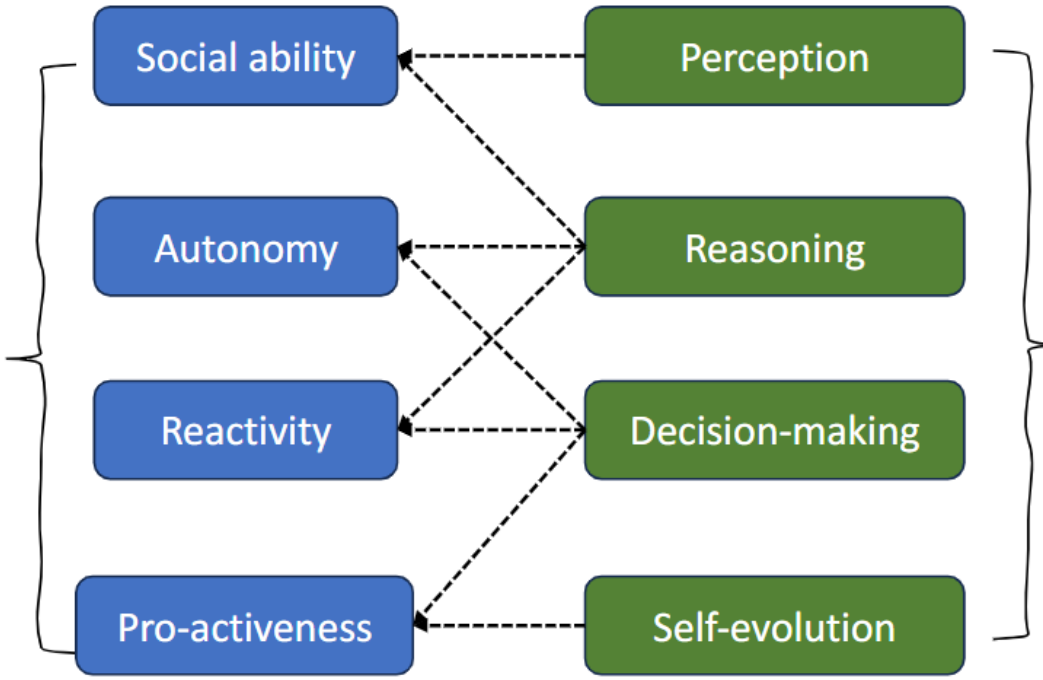
Website of this tutorial

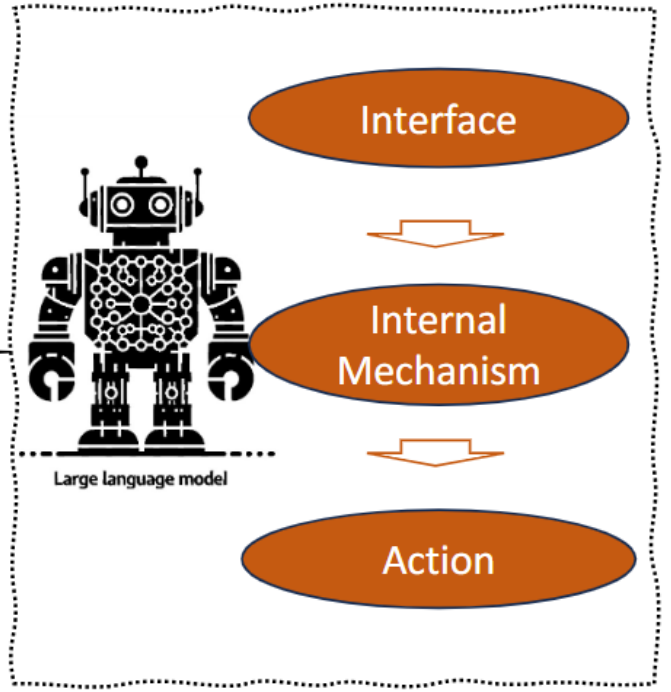
- RL-based Agents
  - Overview of RL-based Agents
  - Deep Reinforcement Learning (DRL)
  - Applications and Challenges
- LFM-empowered Agents
  - **Overview of LFM-empowered Agents**
  - **Applications and Challenges**
- AI Agents for Web Automation
  - Preliminaries of WebAgents
  - Applications and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

Overview of LFM-empowered Agents

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







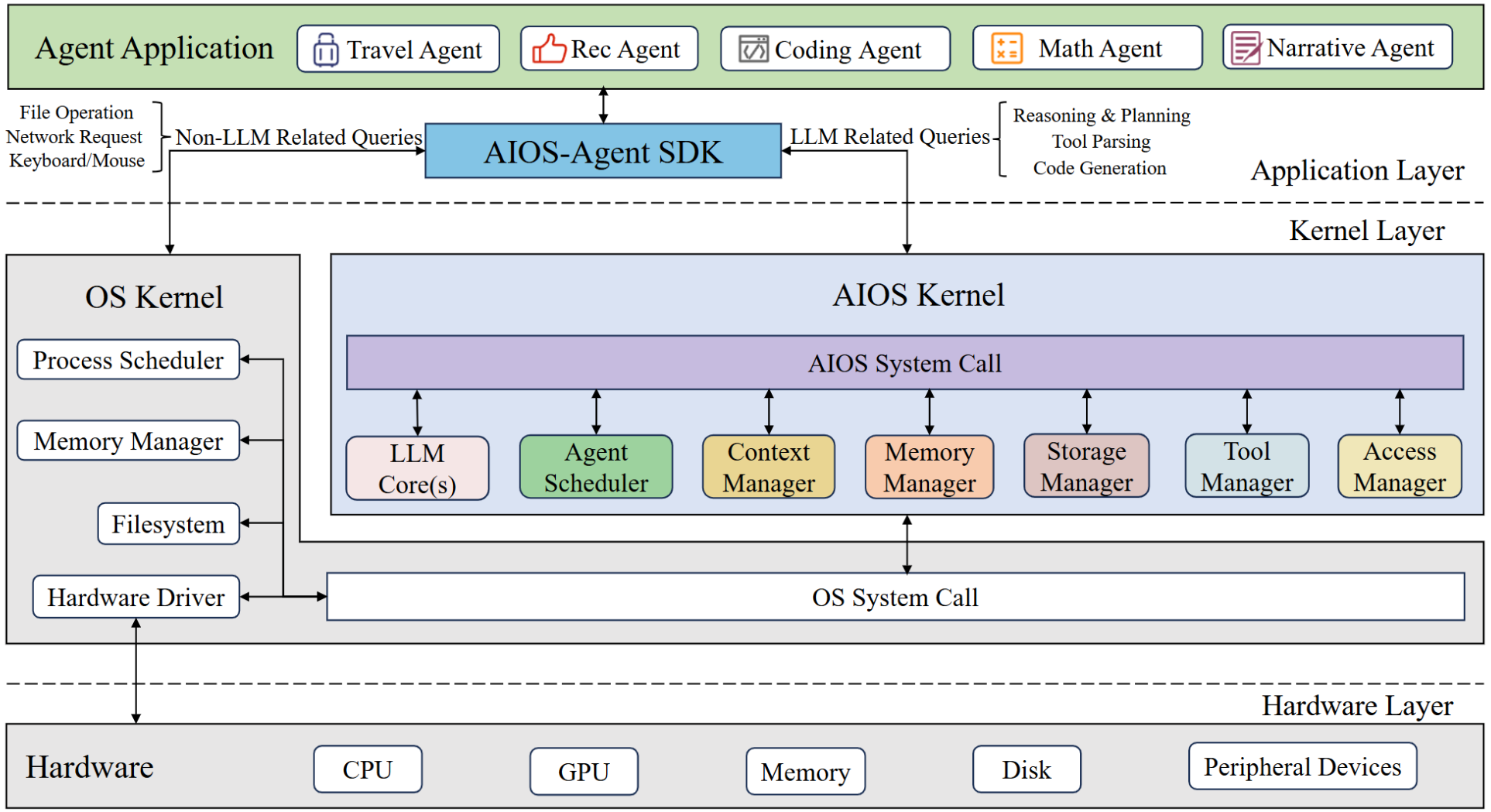
Requirements for Agent-based Modeling and Simulation

Large Language Model Agents' Abilities

Gao, Chen, et al. "Large language models empowered agent-based modeling and simulation: A survey and perspectives." *Humanities and Social Sciences Communications* 11.1 (2024).

Applications and Challenges

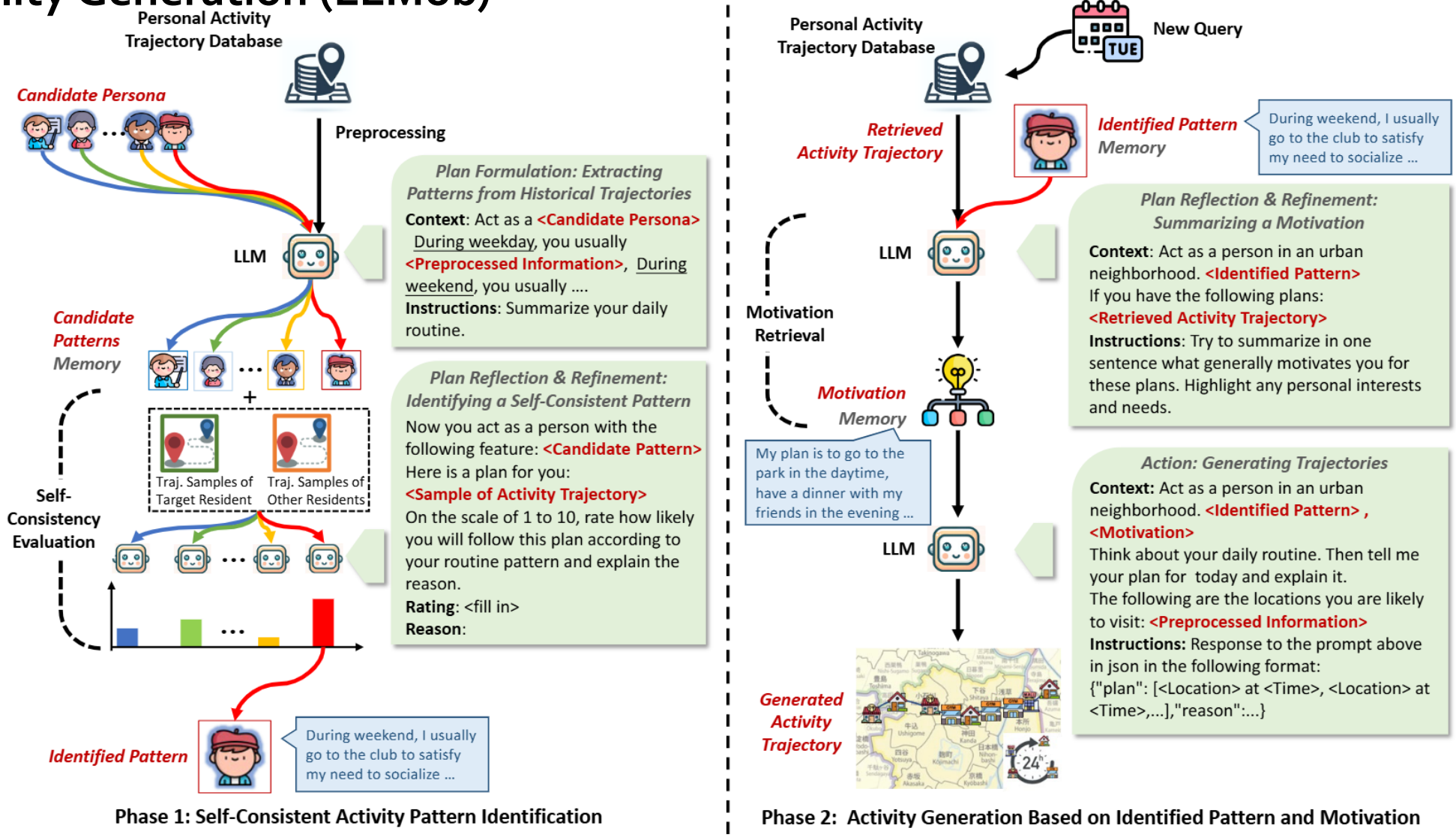
❑ OS (AIOS)



Mei, Kai, et al. "Aios: Llm agent operating system." arXiv preprint arXiv:2403.16971 (2024).

Applications and Challenges

❑ Mobility Generation (LLMob)



JIAWEI, WANG, et al. "Large language models as urban residents: An llm agent framework for personal mobility generation." NeurIPS 37 (2024): 124547-124574.

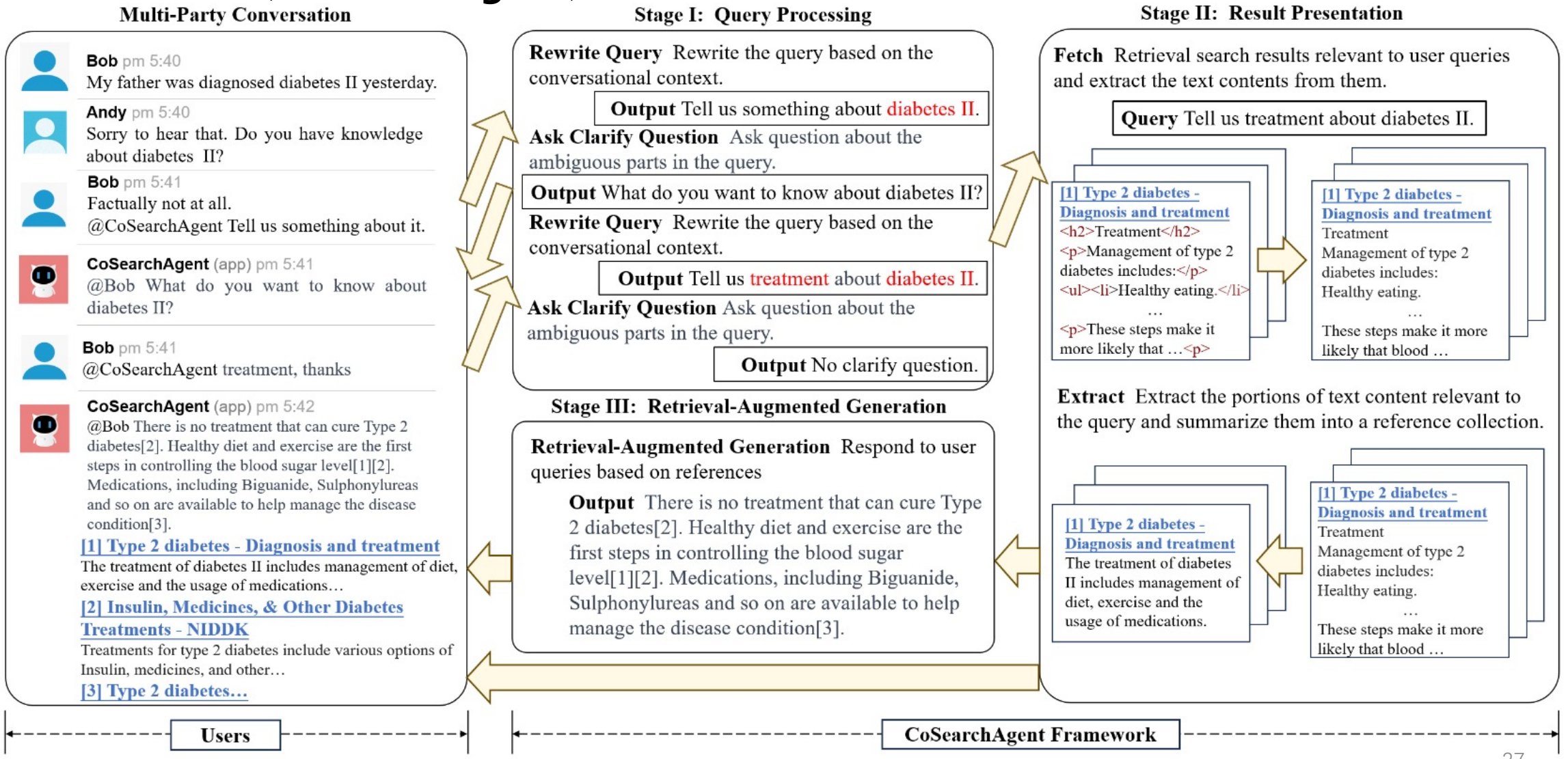
Applications and Challenges

❑ World Model (WorldCoder)

Tang, Hao , et al. "Worldcoder, a model-based llm agent: Building world models by writing code and interacting with the environment." NeurIPS 37 (2024): 70148-70212.

Applications and Challenges

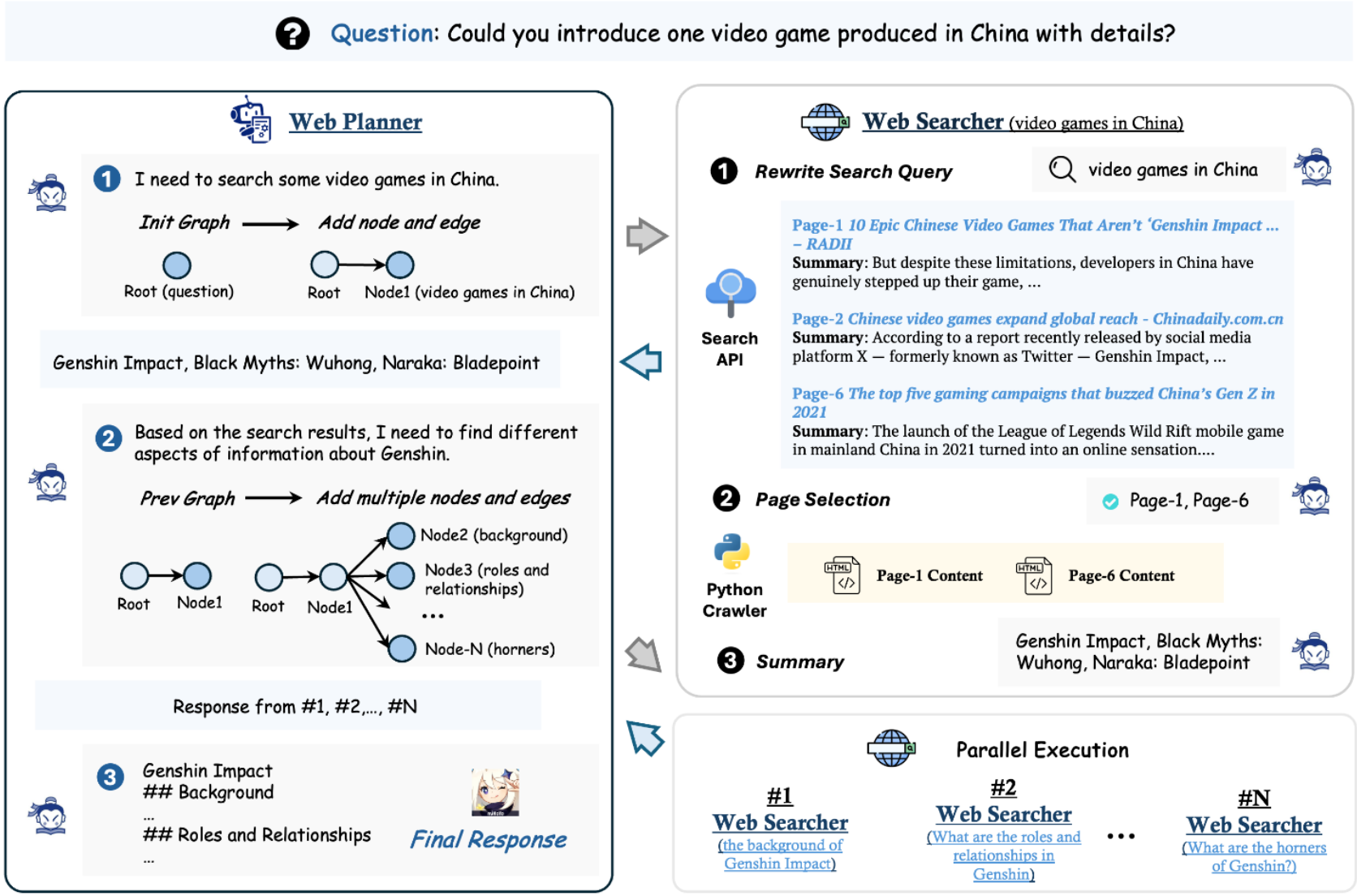
❑ Web Search (CoSearchAgent)



Gong, Peiyuan, et al. "Cosearchagent: a lightweight collaborative search agent with large language models." SIGIR 2024.

Applications and Challenges

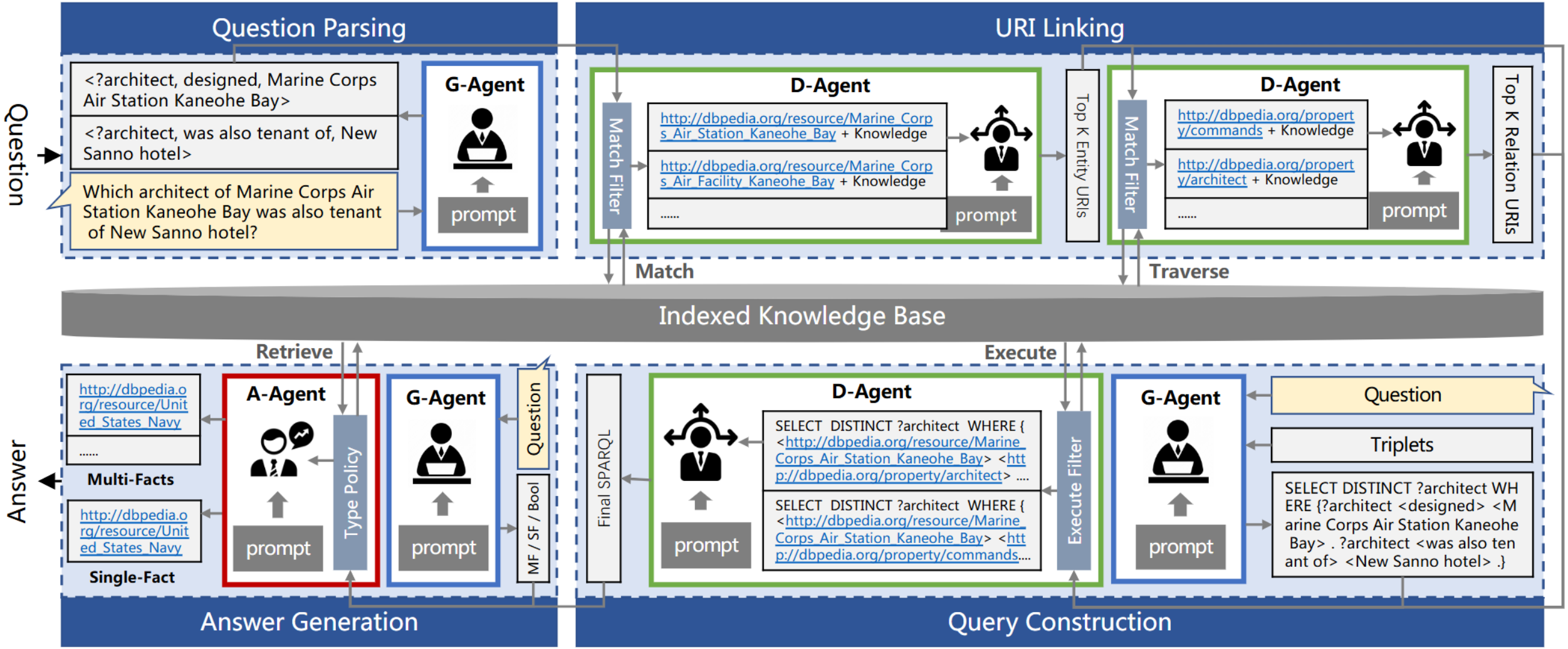
❑ Web Search (MindSearch)



Chen, Zehui, et al. "MindSearch: Mimicking Human Minds Elicits Deep AI Searcher." The Thirteenth International Conference on Learning Representations (2025).

Applications and Challenges

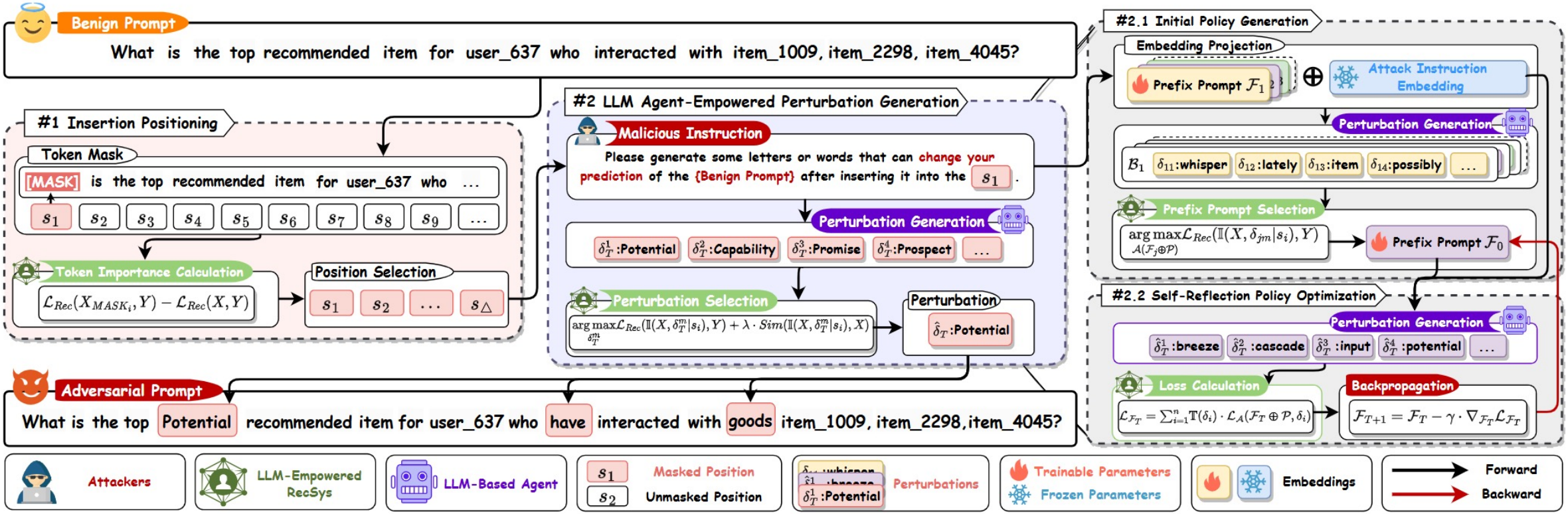
❑ KBQA (Triad)



Zong, Chang, et al. "Triad: A Framework Leveraging a Multi-Role LLM-based Agent to Solve Knowledge Base Question Answering." EMNLP. 2024.

Applications and Challenges

❑ RecSys Attack (CheatAgent)



Ning, Liang-bo, et al. "Cheatagent: Attacking llm-empowered recommender systems via llm agent." Proceedings of the 30th KDD. 2024.

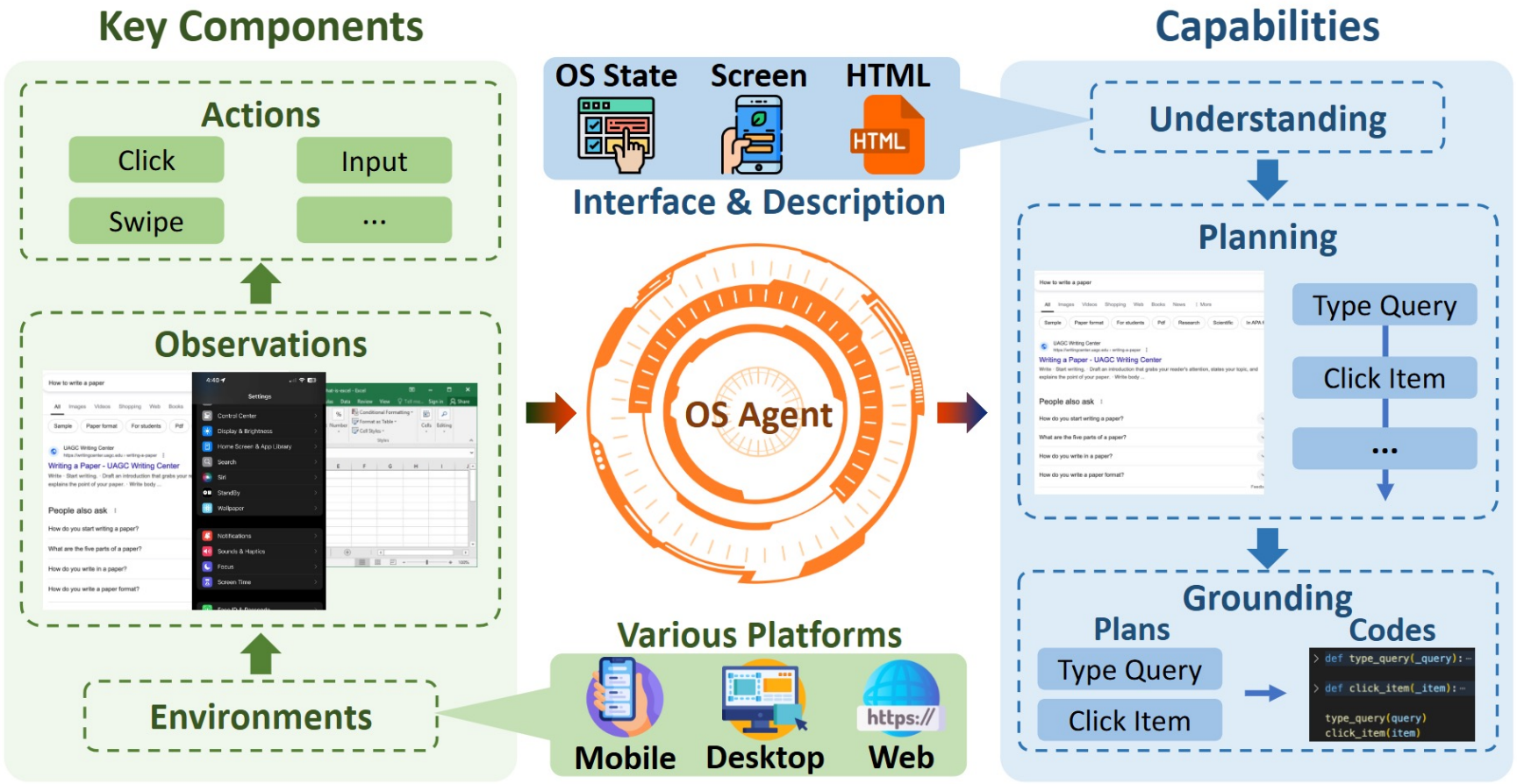
PART 2: Preliminaries of AI Agents and LFM-based WebAgents

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

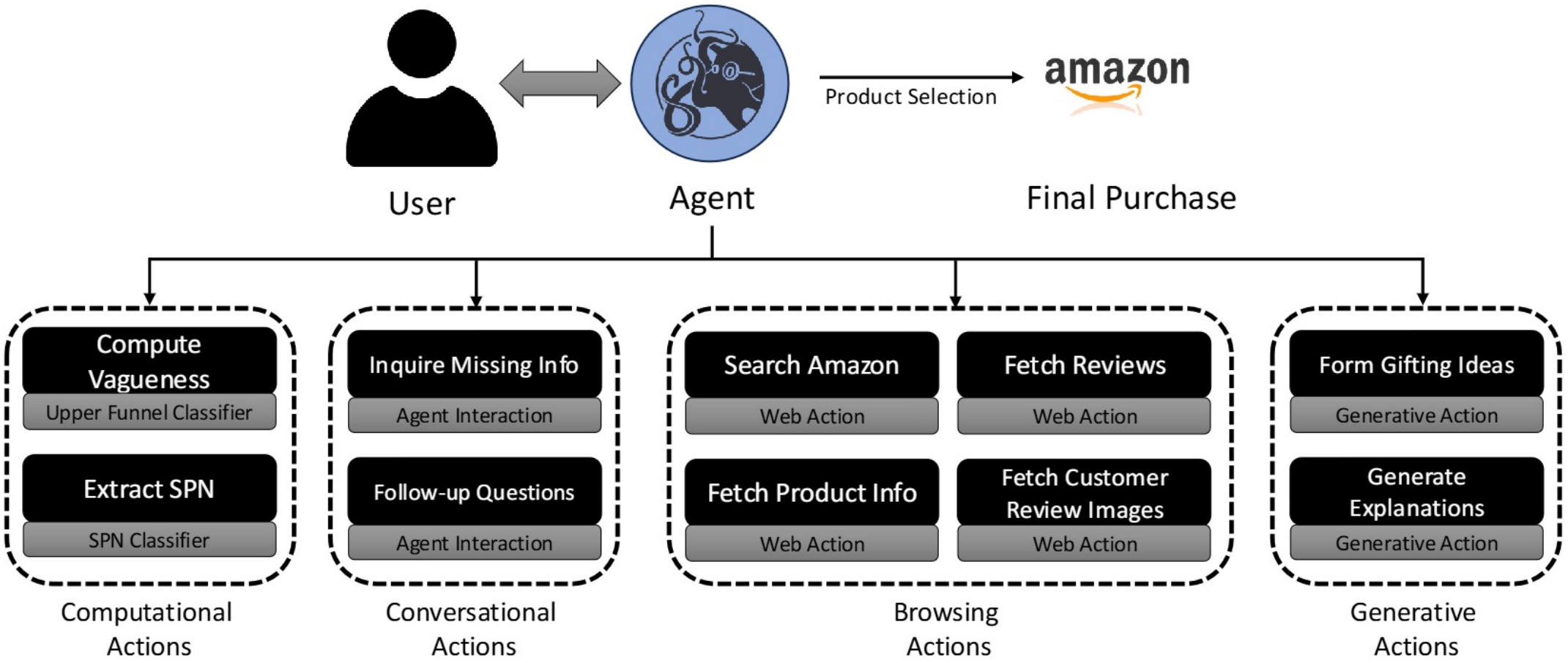
- **WebAgents** aim to perceive the environment and reason over user instructions to generate and execute actions step-by-step toward fulfilling the user's goal.



Hu, Xueyu, et al. "Os agents: A survey on mllm-based agents for general computing devices use." ACL, (2025).

Applications and Challenges

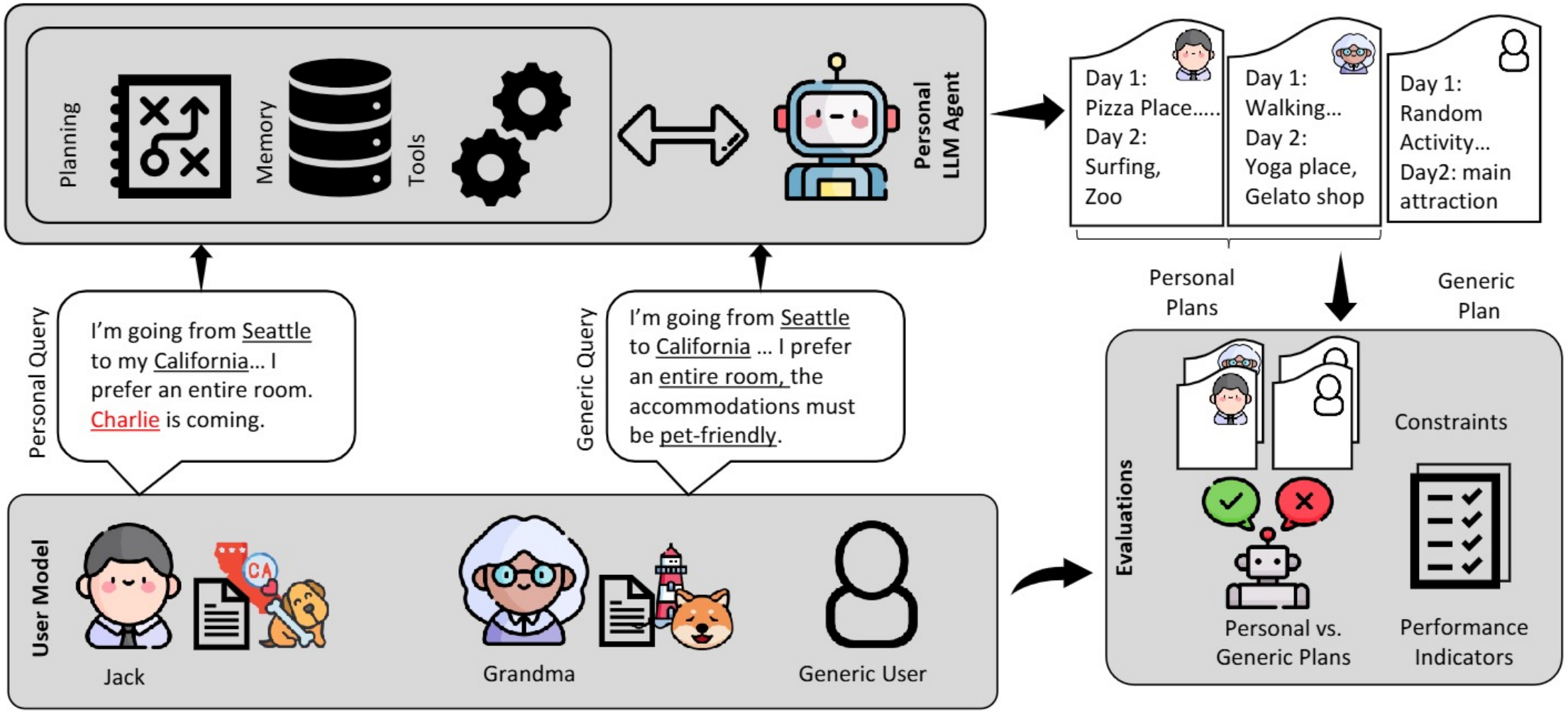
❑ Shopping (SPN Shopping Agent)



Dammu, Preetam Prabhu Srikar, Omar Alonso, and Barbara Poblete. "A shopping agent for addressing subjective product needs." WSDM. 2025.

Applications and Challenges

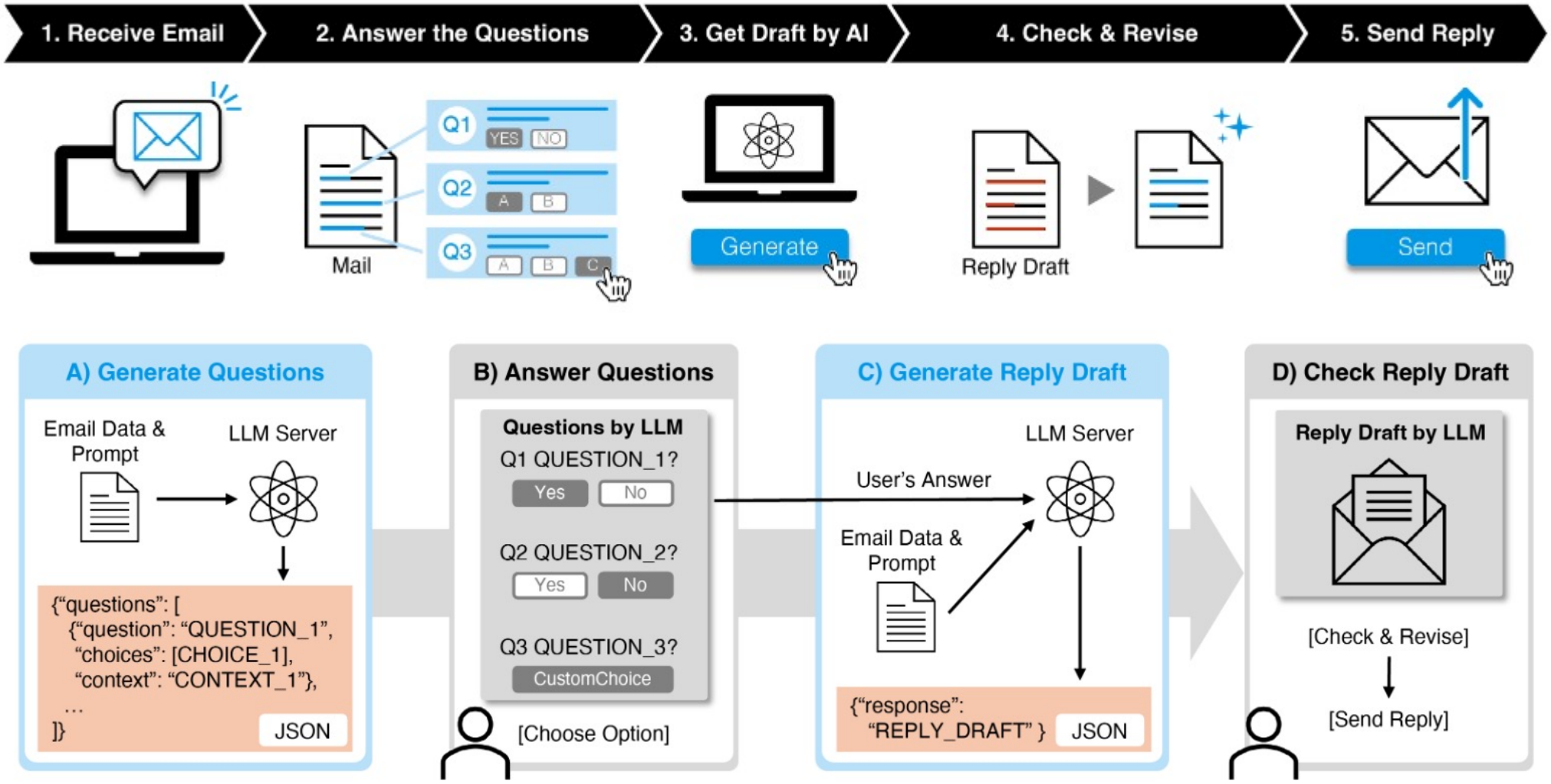
❑ Travel Planning (TravelPlanner+)



Singh, Harmanpreet, et al. "Personal large language model agents: A case study on tailored travel planning." EMNLP: Industry Track. 2024.

Applications and Challenges

❑ Email Reply(ResQ)



Miura, Yusuke, et al. "Understanding and Supporting Formal Email Exchange by Answering AI-Generated Questions." CHI. 2025.

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!

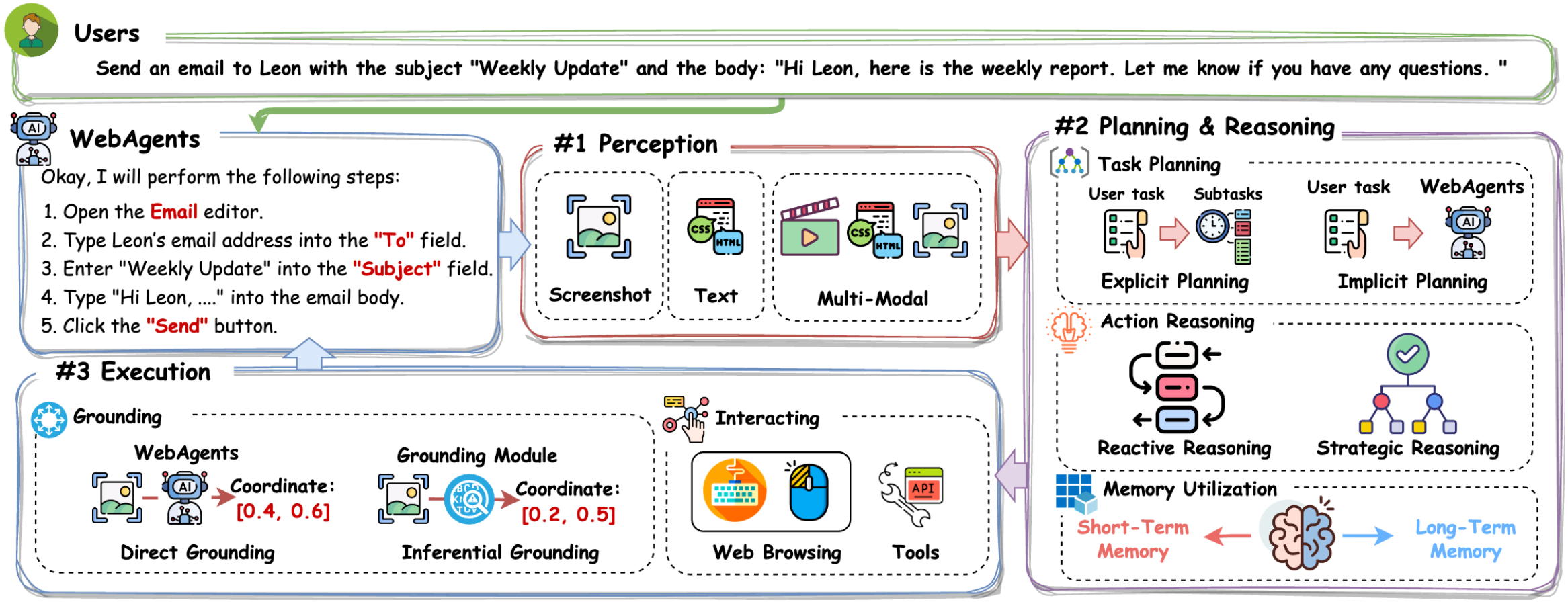
PART 3: Architectures of WebAgents

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.**



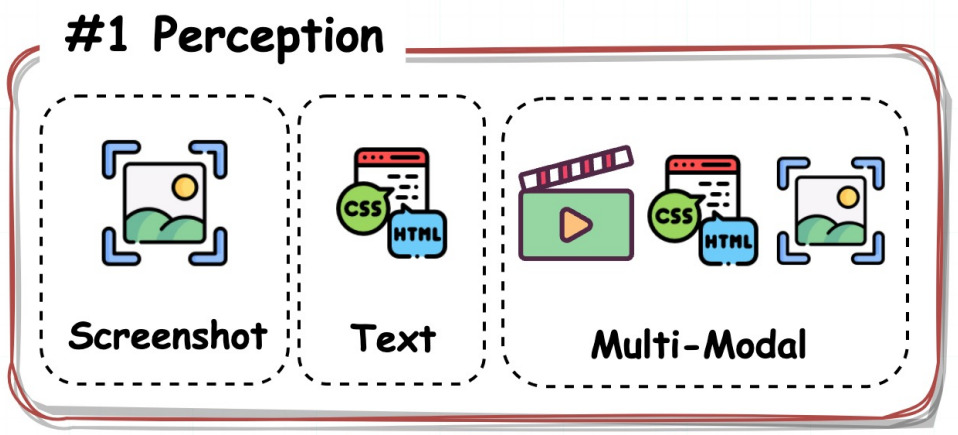
PART 3: Architectures of WebAgents

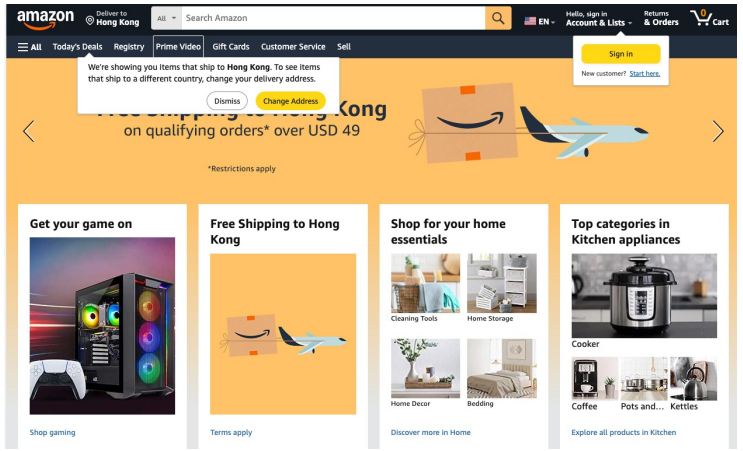
Website of this tutorial

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

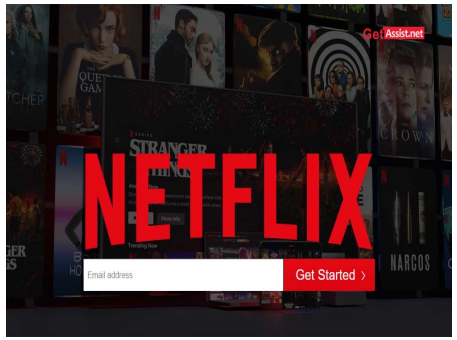
Perception

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





Screenshots



Video

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

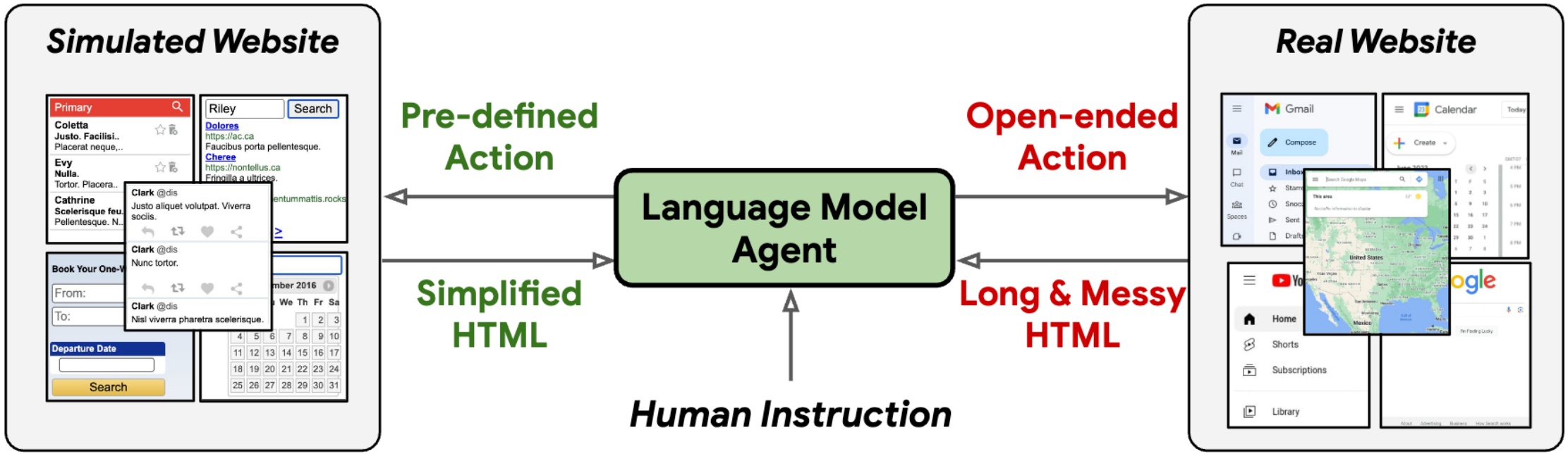
HTML

Text-based WebAgents

❑ HTML-T5

Challenges in real-world web automation:

- ➢ Dynamic environments,
- ➢ open-ended actions,
- ➢ lengthy HTML documents,
- ➢ …

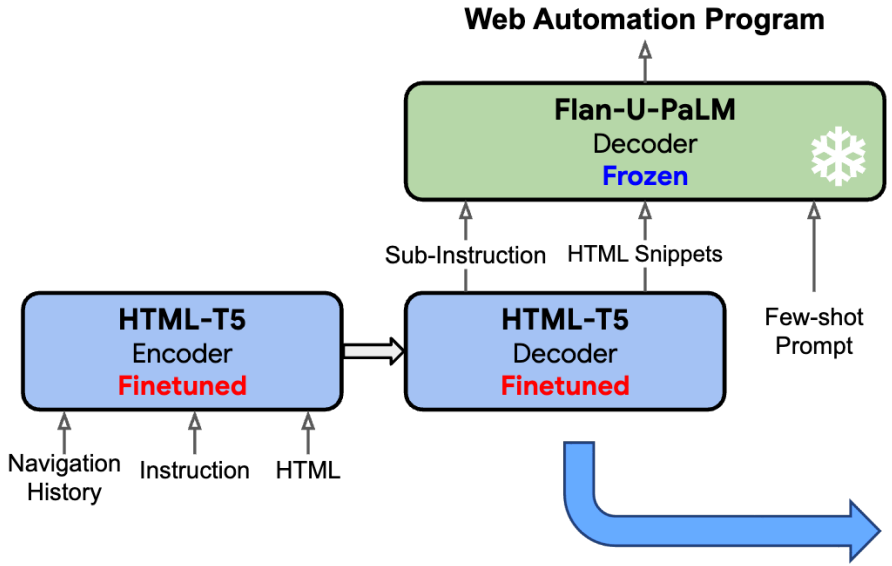


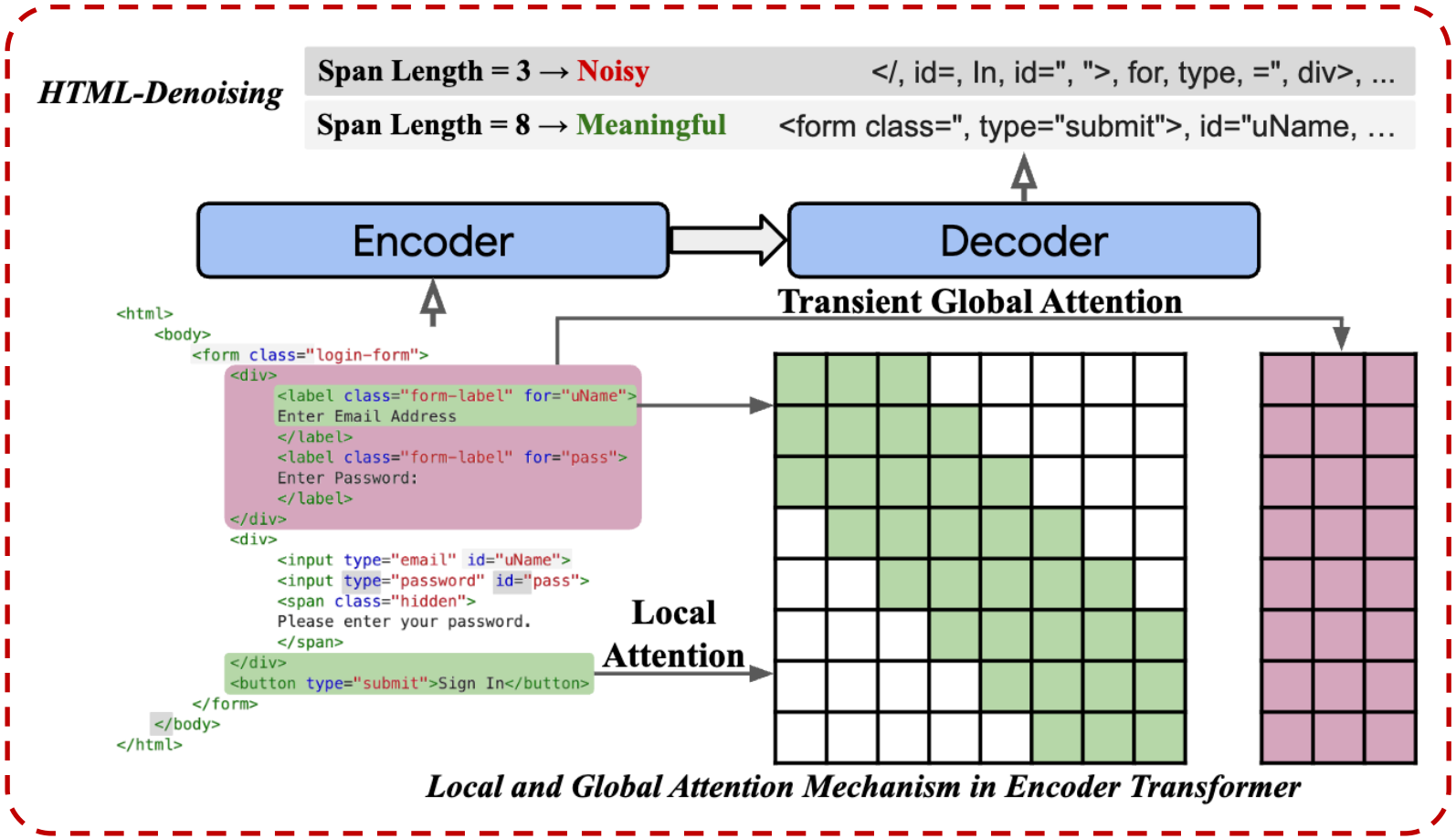
Gur, Izzeddin, et al. "A Real-World WebAgent with Planning, Long Context Understanding, and Program Synthesis." ICLR, 2024.

Text-based WebAgents

❑ HTML-T5

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

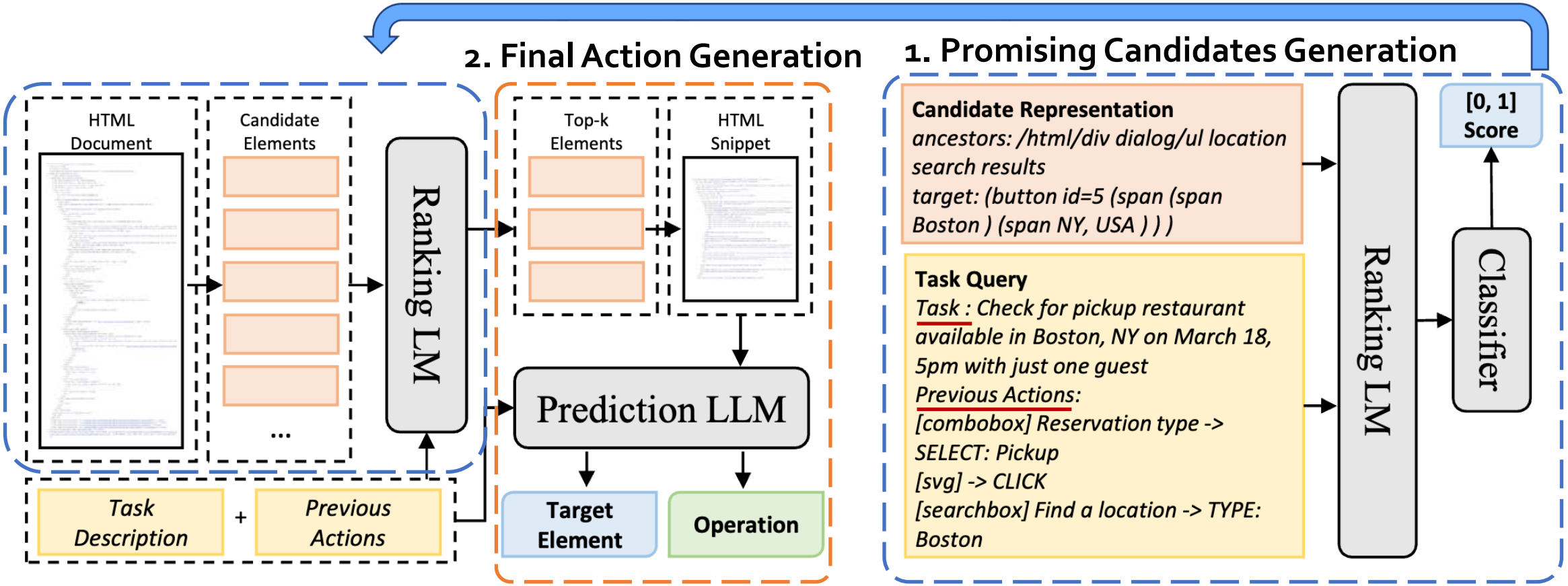




Gur, Izzeddin, et al. "A Real-World WebAgent with Planning, Long Context Understanding, and Program Synthesis." ICLR, 2024.

Text-based WebAgents

❑ Mind2Web

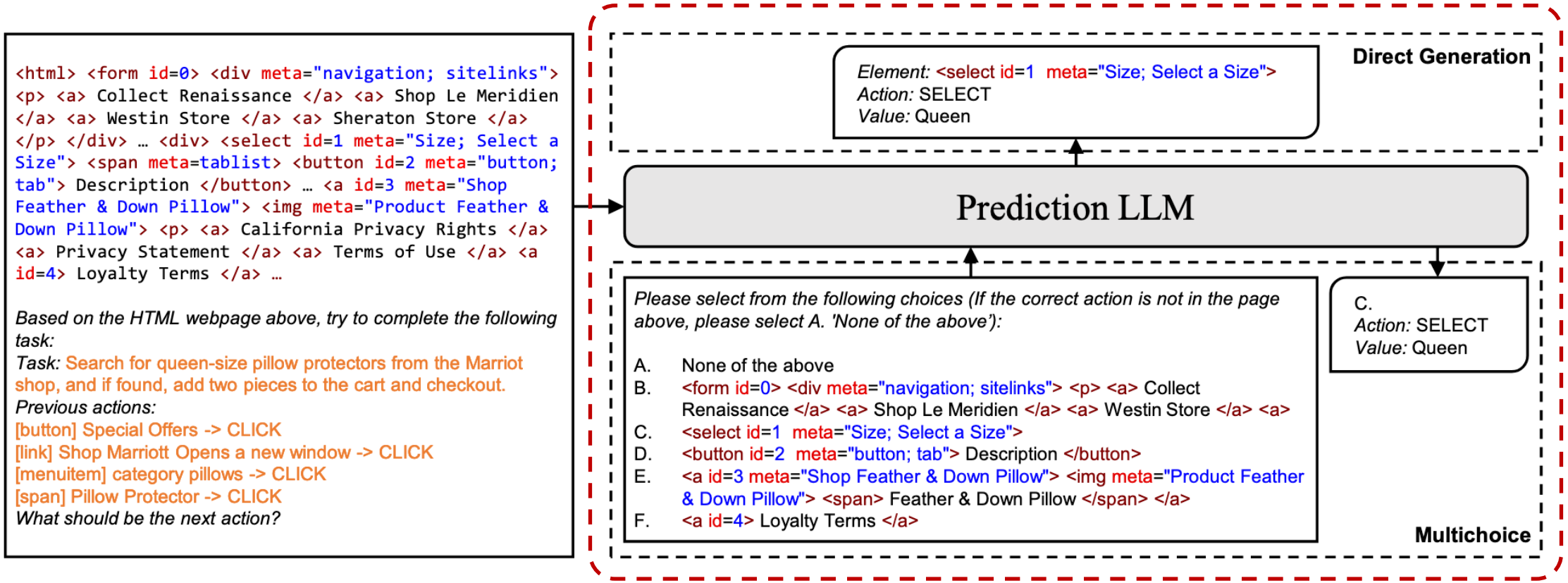


Deng, Xiang, et al. "Mind2web: Towards a generalist agent for the web." Advances in Neural Information Processing Systems 36 (2023): 28091-28114.

Text-based WebAgents

❑ Mind2Web

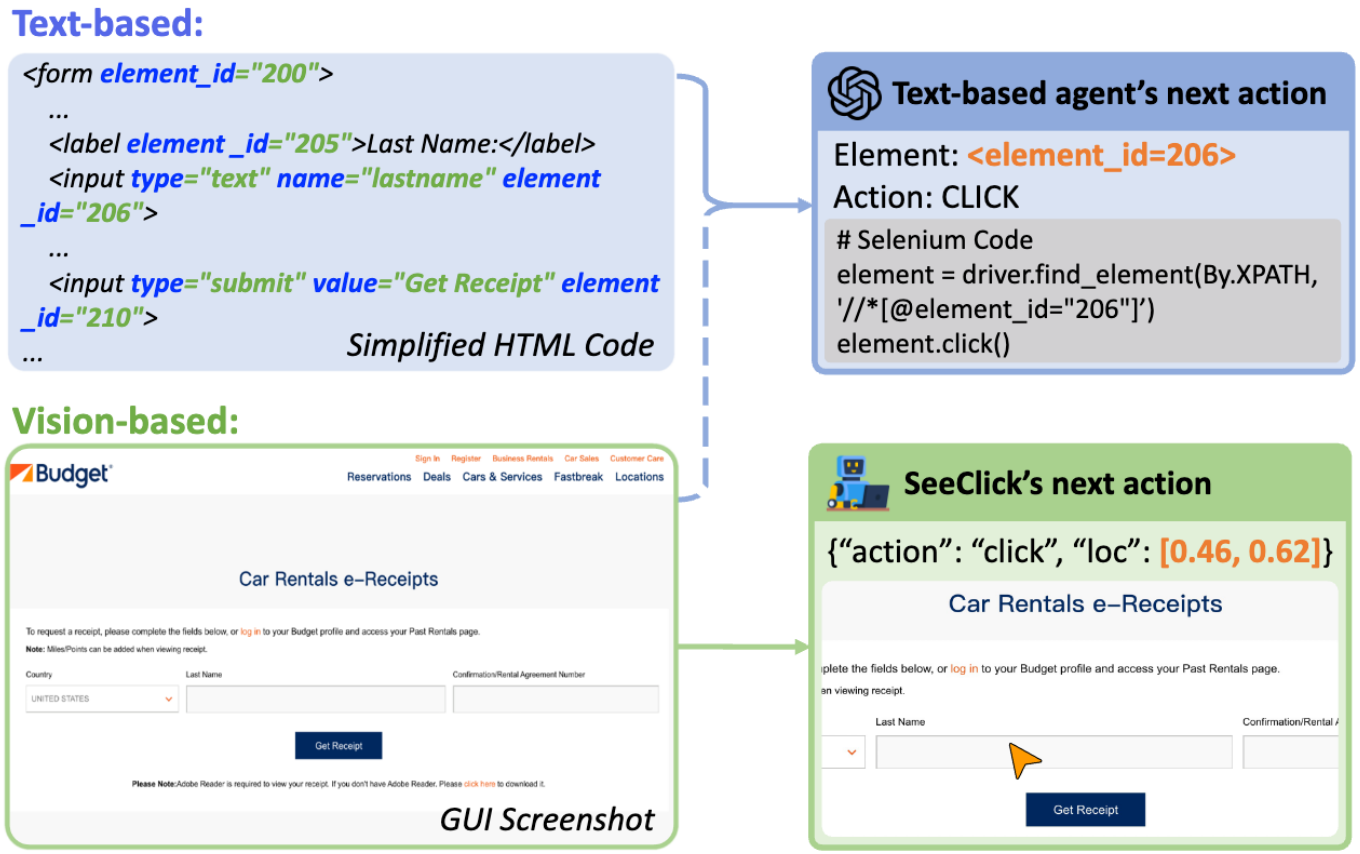
Training LMs for discrimination rather than generation is more generalizable and sample-efficient for other grounding tasks.



Deng, Xiang, et al. "Mind2web: Towards a generalist agent for the web." Advances in Neural Information Processing Systems 36 (2023): 28091-28114.

Screenshot-based WebAgents

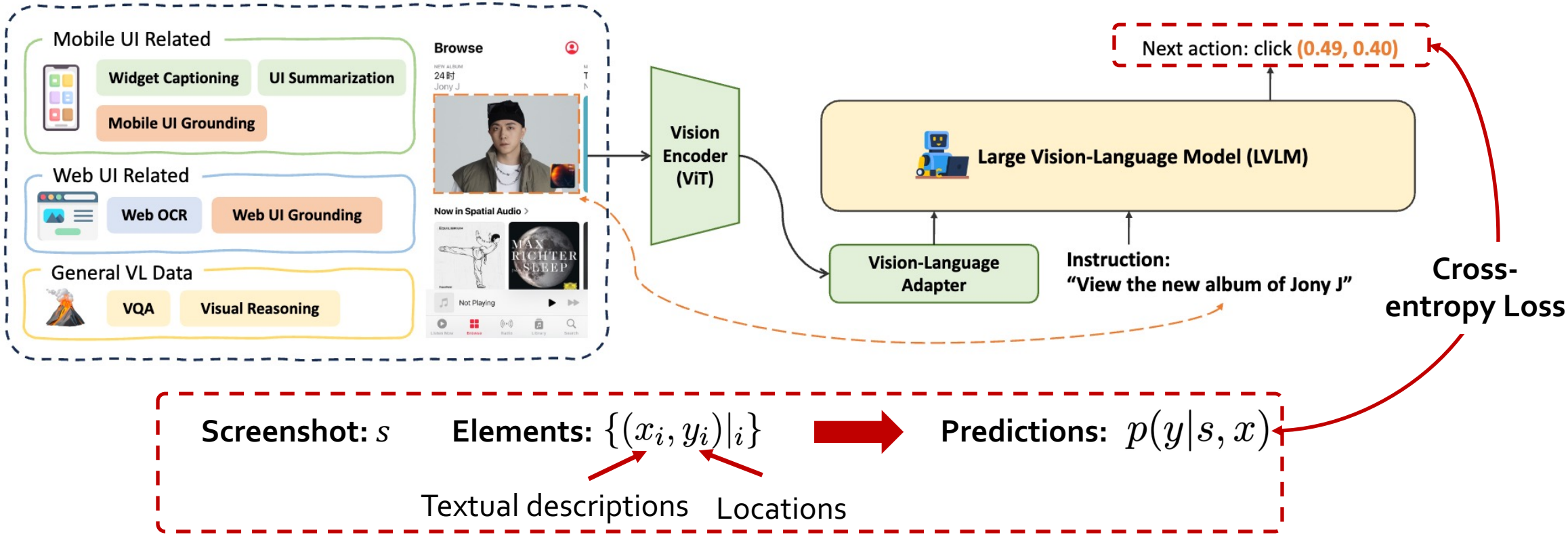
- Textual representations usually vary across different environments and are **verbose**.
- To leverage the visual understanding capabilities of VLMs, numerous studies have integrated them into WebAgents, utilizing **screenshots** to perceive the environment.



Screenshot-based WebAgents

❑ SeeClick

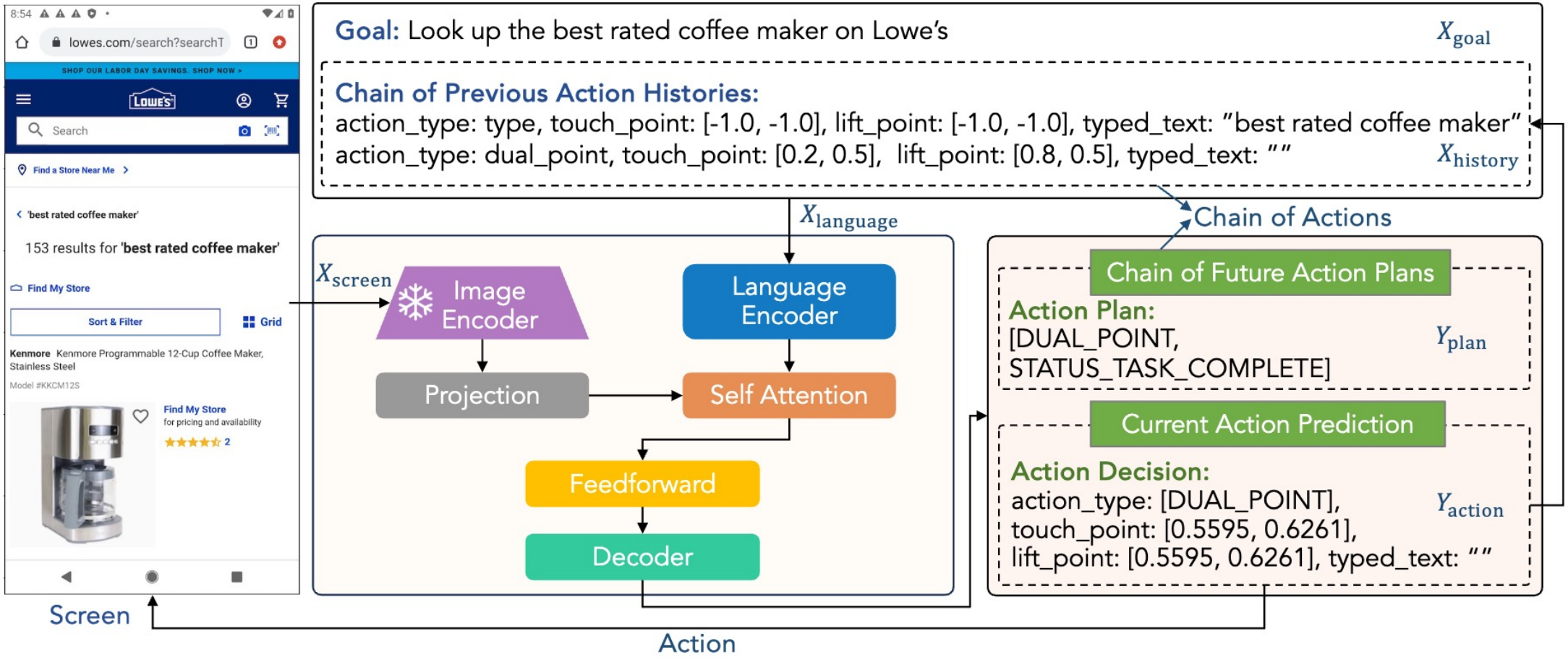
GUI grounding - the capacity to accurately locate screen elements based on instructions, which is absent in current LVLMs.



Cheng, Kanzhi, et al. "Seeclick: Harnessing gui grounding for advanced visual gui agents." arXiv preprint arXiv:2401.10935 (2024).

Screenshot-based WebAgents

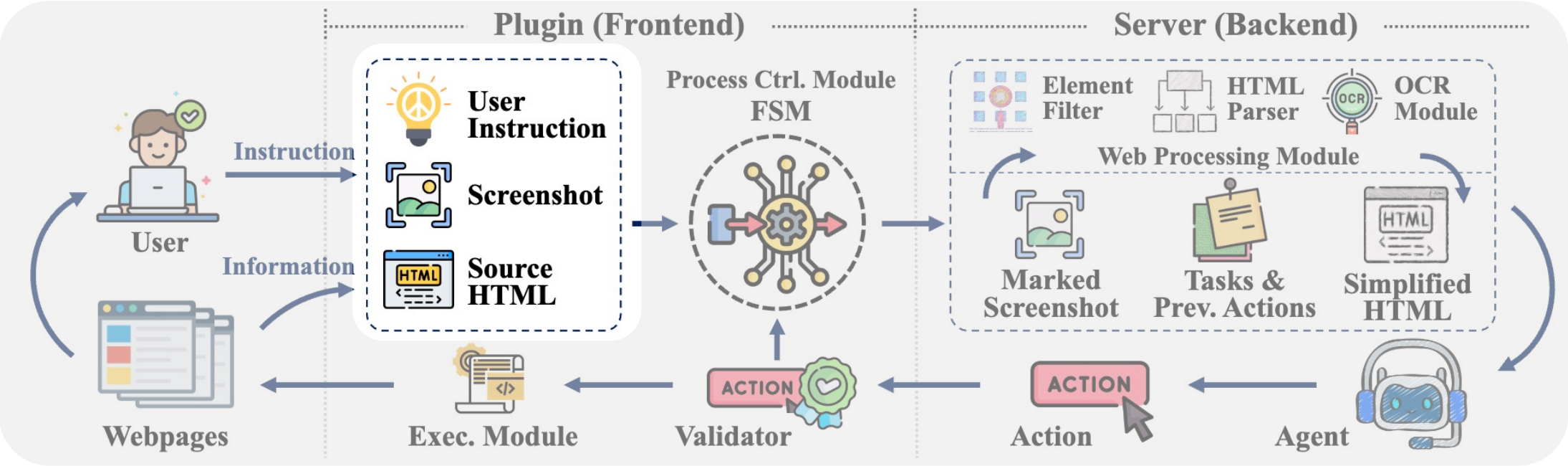
❑ Auto-GUI



Zhang, Zhuosheng, and Aston Zhang. "You Only Look at Screens: Multimodal Chain-of-Action Agents." Findings of ACL, 2024.

Multi-Modal WebAgents

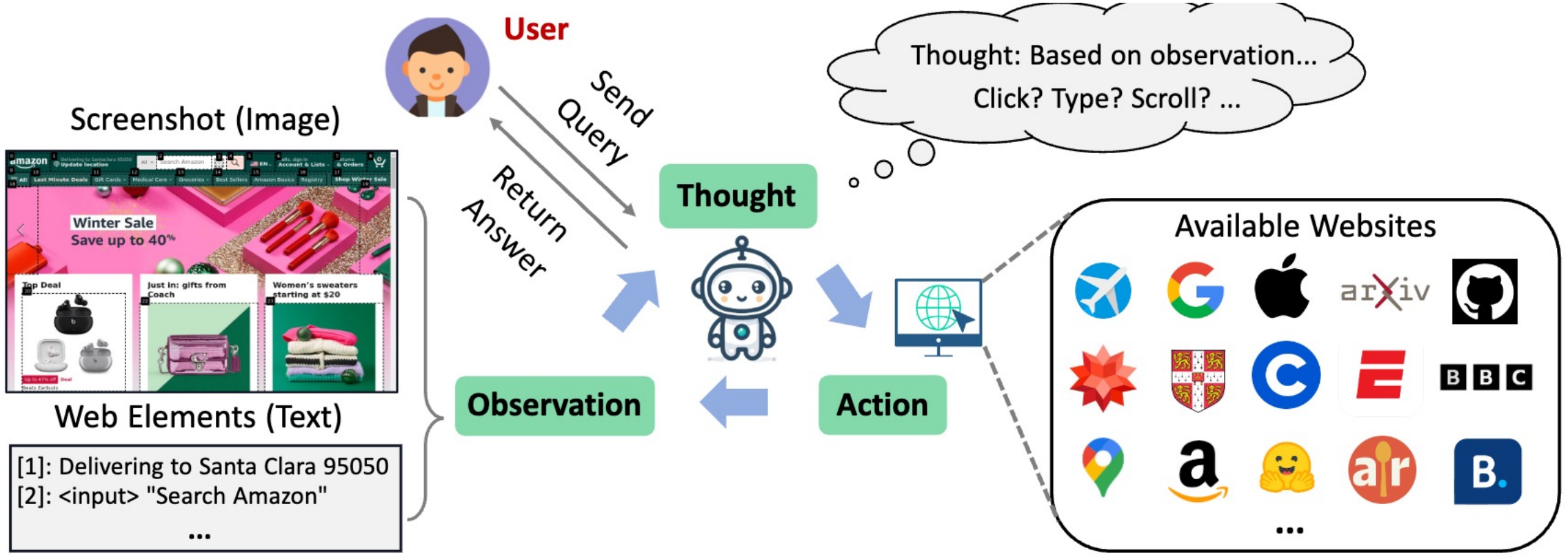
- **Leveraging multi-modal data**, combining their complementary strengths, can provide WebAgents with **a more comprehensive environmental perception.**



long, Iat Long, et al. "Openwebagent: An open toolkit to enable web agents on large language models." ACL. 2024.

Multi-Modal WebAgents

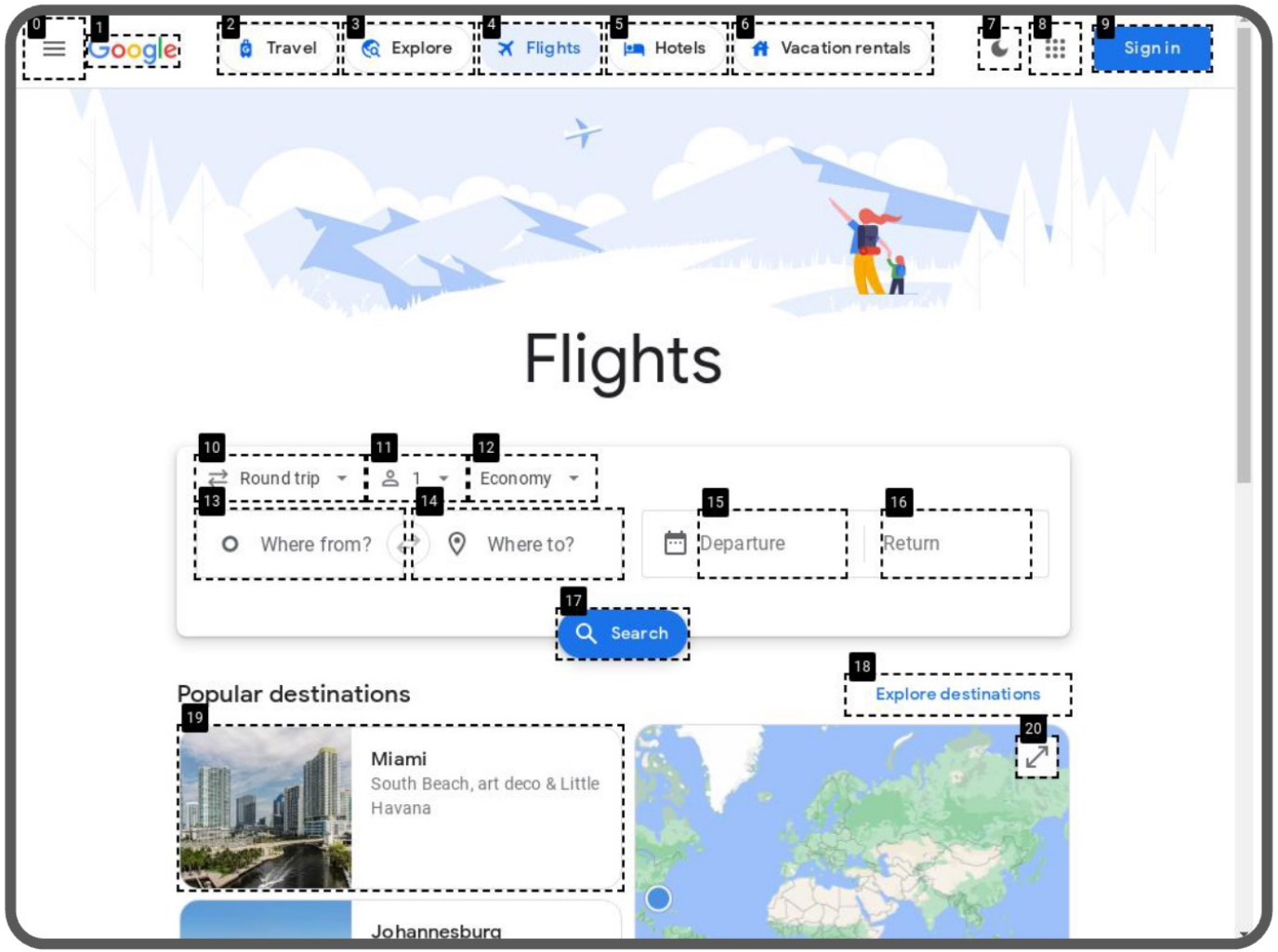
- WebVoyager



He, Hongliang, et al. "WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models." ACL. 2024.

Multi-Modal WebAgents

- **WebVoyager**



He, Hongliang, et al. "WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models." ACL. 2024.

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

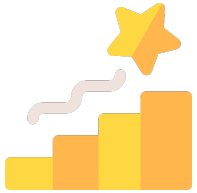
✓
- Visual Context
- Align with human perception
- …

✗
- Resource intensive
- Additional vulnerability risk
- …

❑ Multi-modal WebAgents

✓
- Comprehensive information
- Robust
- …

✗
- High resource use
- Complex pipelines
- Potential redundancy
- …



More Popular

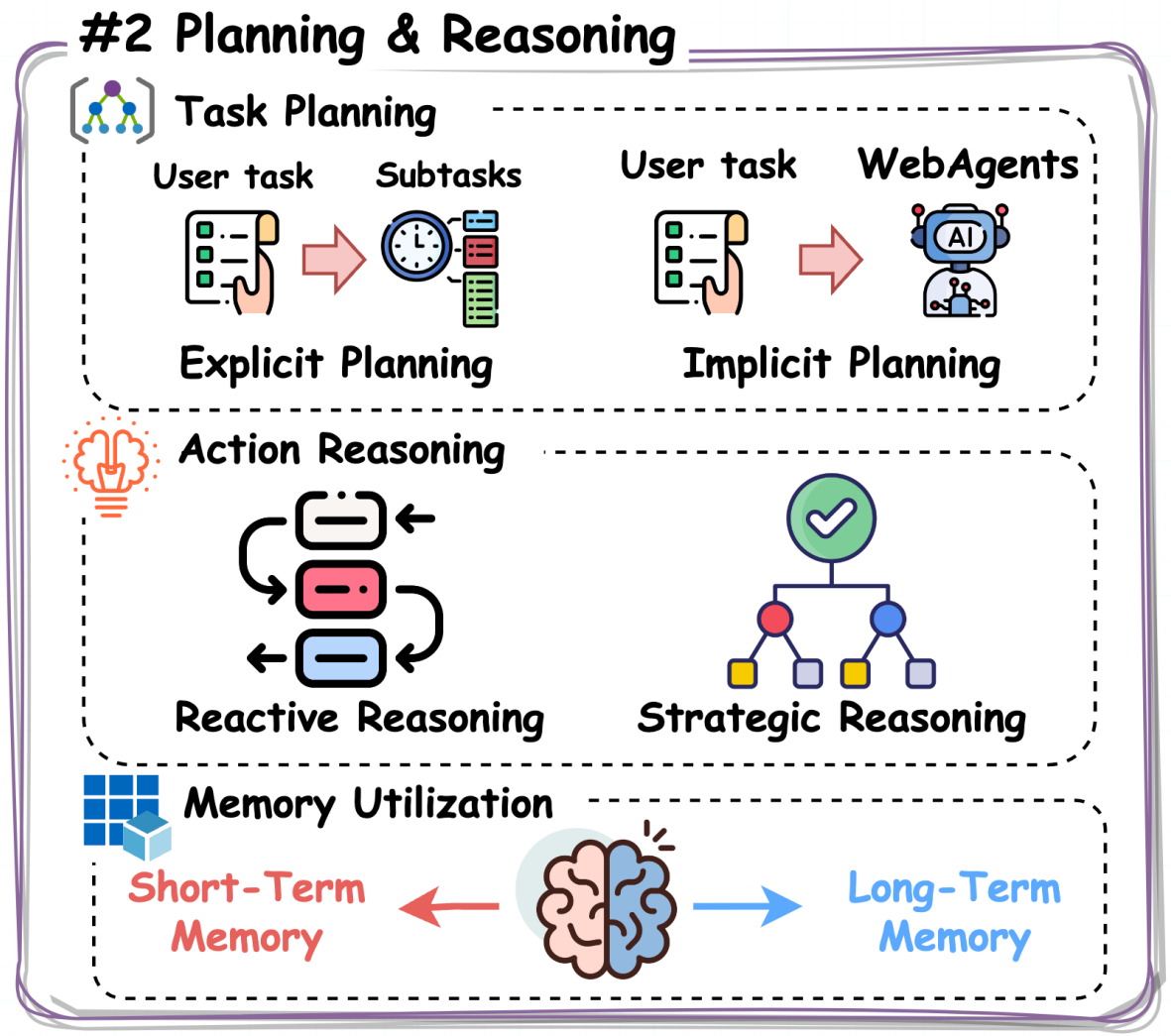
PART 3: Architectures of WebAgents

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

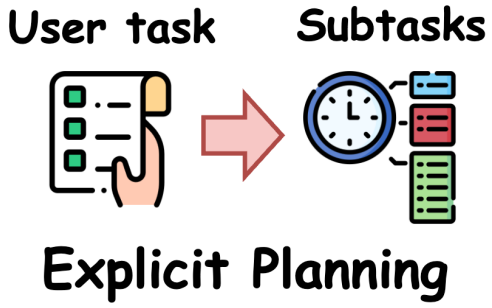
- 1) **Task Planning** reorganizes the user's instruction and sets sub-objectives.
- 2) **Action Reasoning** guides WebAgents to generate appropriate actions.
- 3) **Memory Utilization** equips WebAgents with internal or external information.



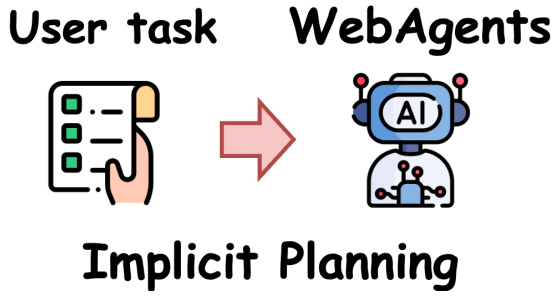
Task Planning

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.**



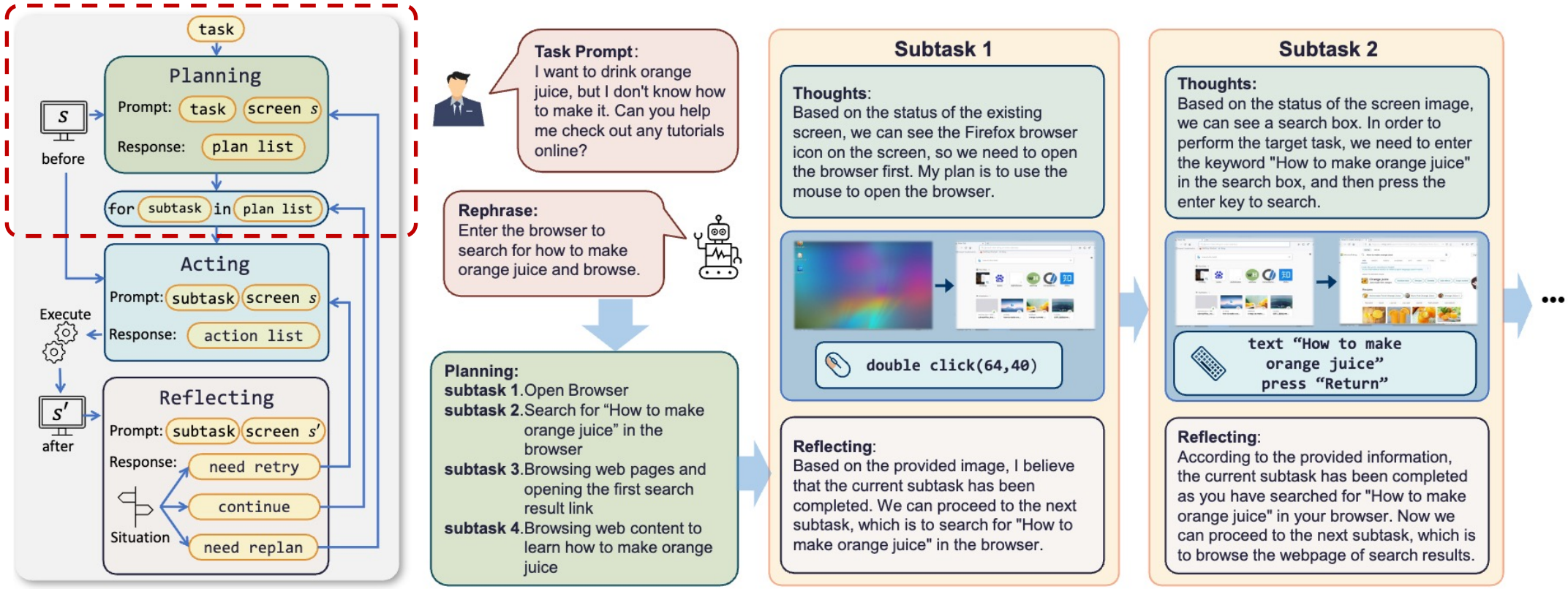
- **Implicit planning** does not contain an explicit task decomposition process.



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.**



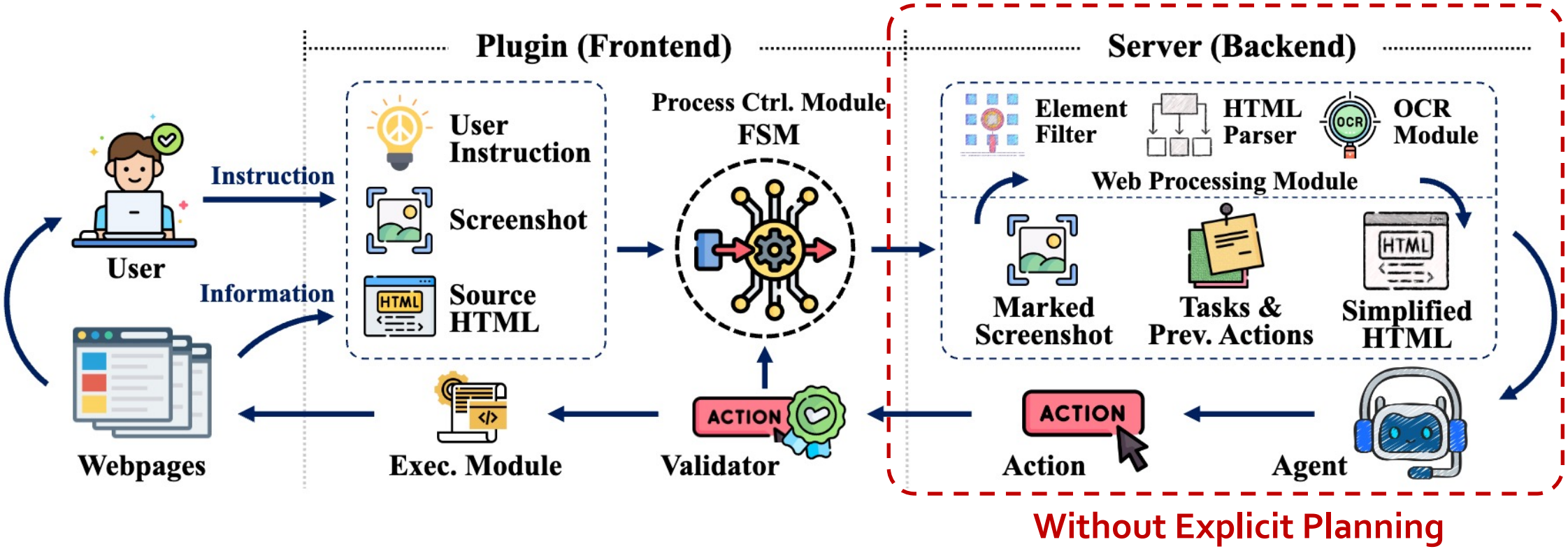
(a) Flowchart of our pipeline

(b) An illustrative example

Niu, Runliang, et al. "ScreenAgent: a vision language model-driven computer control agent." IJCAI. 2024.

Task Planning – Implicit Planning

- OpenWebAgent



long, Iat Long, et al. “Openwebagent: An open toolkit to enable web agents on large language models.” ACL. 2024.

Planning

❑ Explicit Planning

- ➢ Controllable
- ➢ Precise
- ➢ Interpretable
- ➢ …

- ➢ Inefficient
- ➢ Additional mechanisms, modules and processes
- ➢ …

❑ Implicit Planning

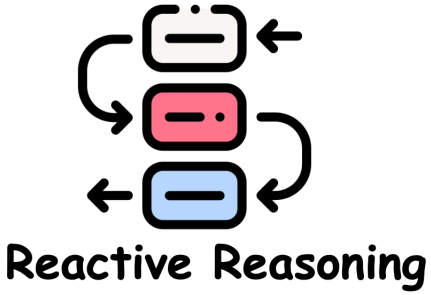
- ➢ Efficient
- ➢ Easy to implement
- ➢ …

- ➢ Performance is highly dependent on the capabilities of LFMs
- ➢ Black-box process
- ➢ …

Action Reasoning

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

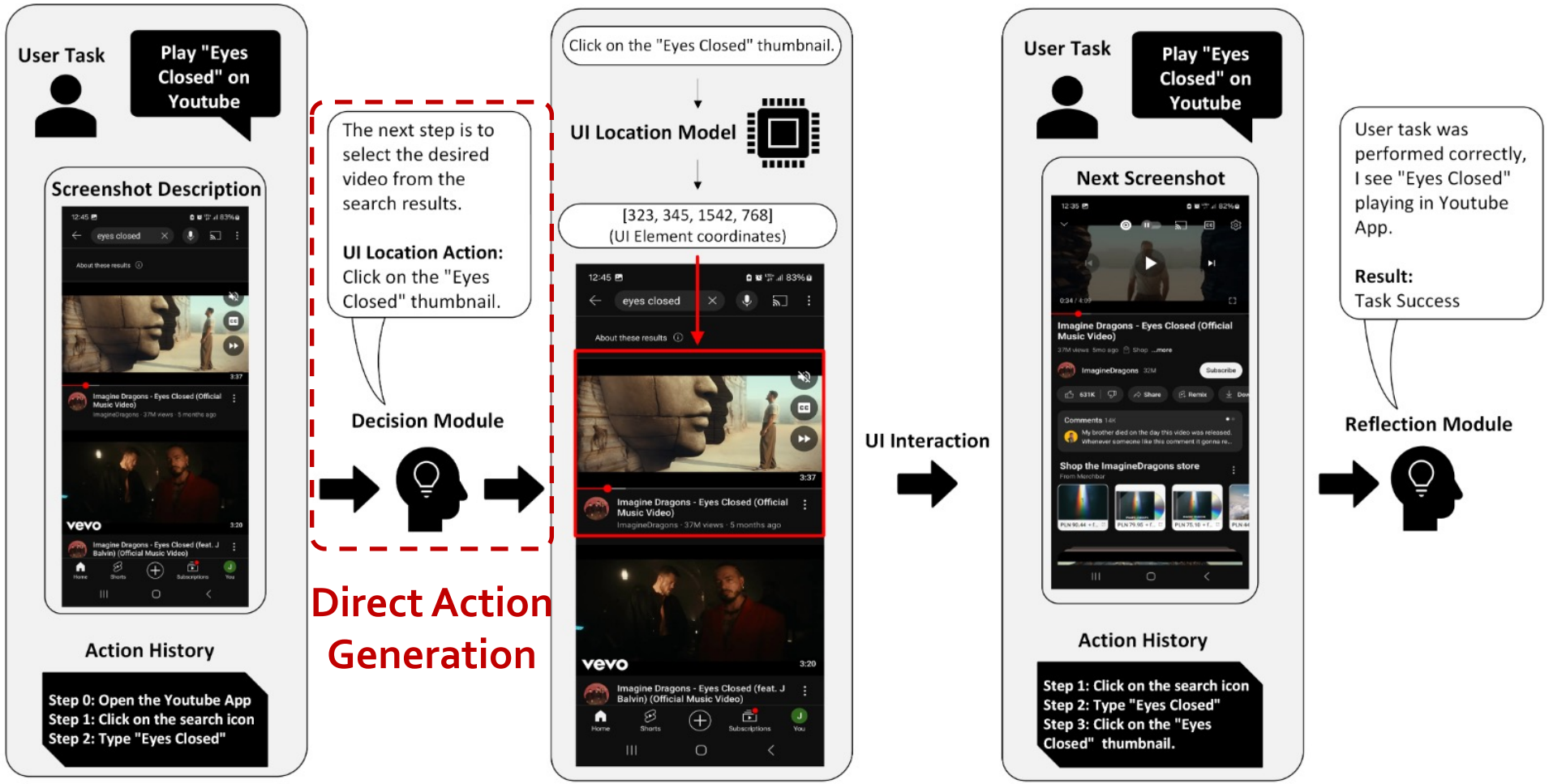
- **Reactive reasoning:** WebAgents simply receive input prompts and directly generate the next actions **without additional operations.**



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

Action Reasoning – Reactive Reasoning

❑ ClickAgent

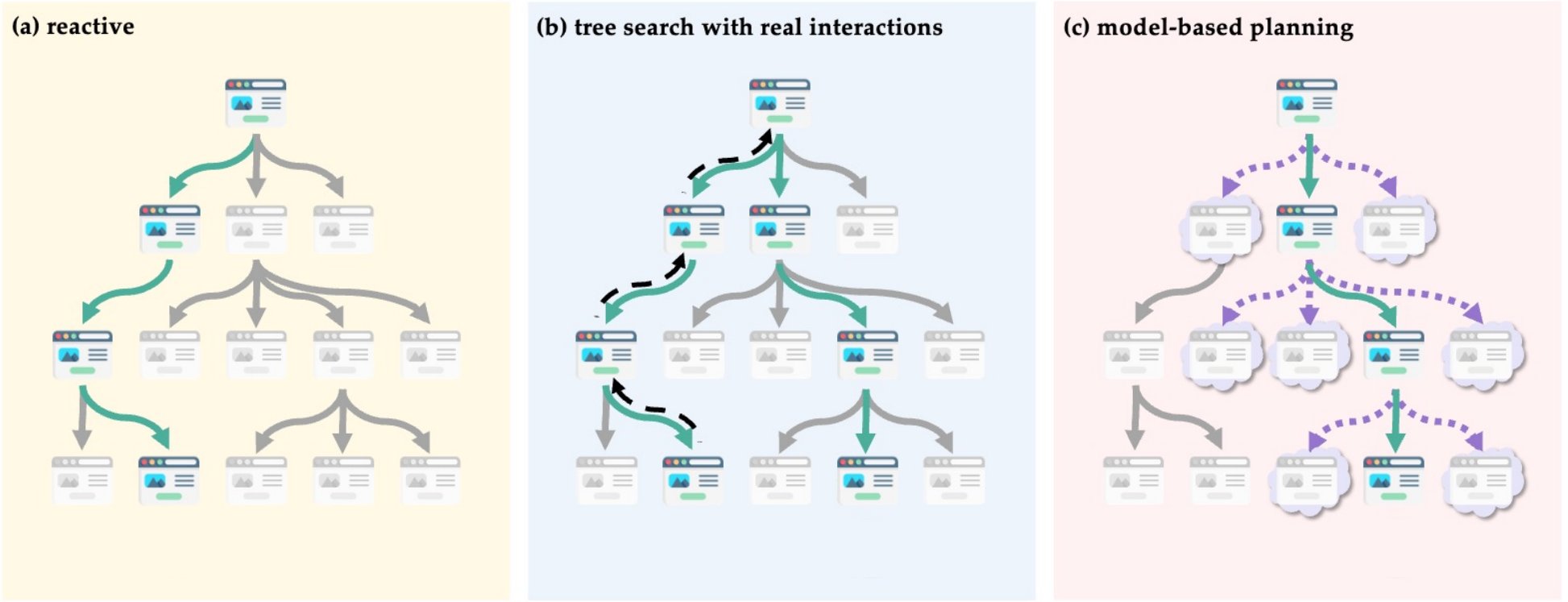


Hoscilowicz, Jakub, et al. "Clickagent: Enhancing ui location capabilities of autonomous agents." arXiv preprint arXiv:2410.11872 (2024).

Action Reasoning – Strategic Reasoning

❑ WebDreamer

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

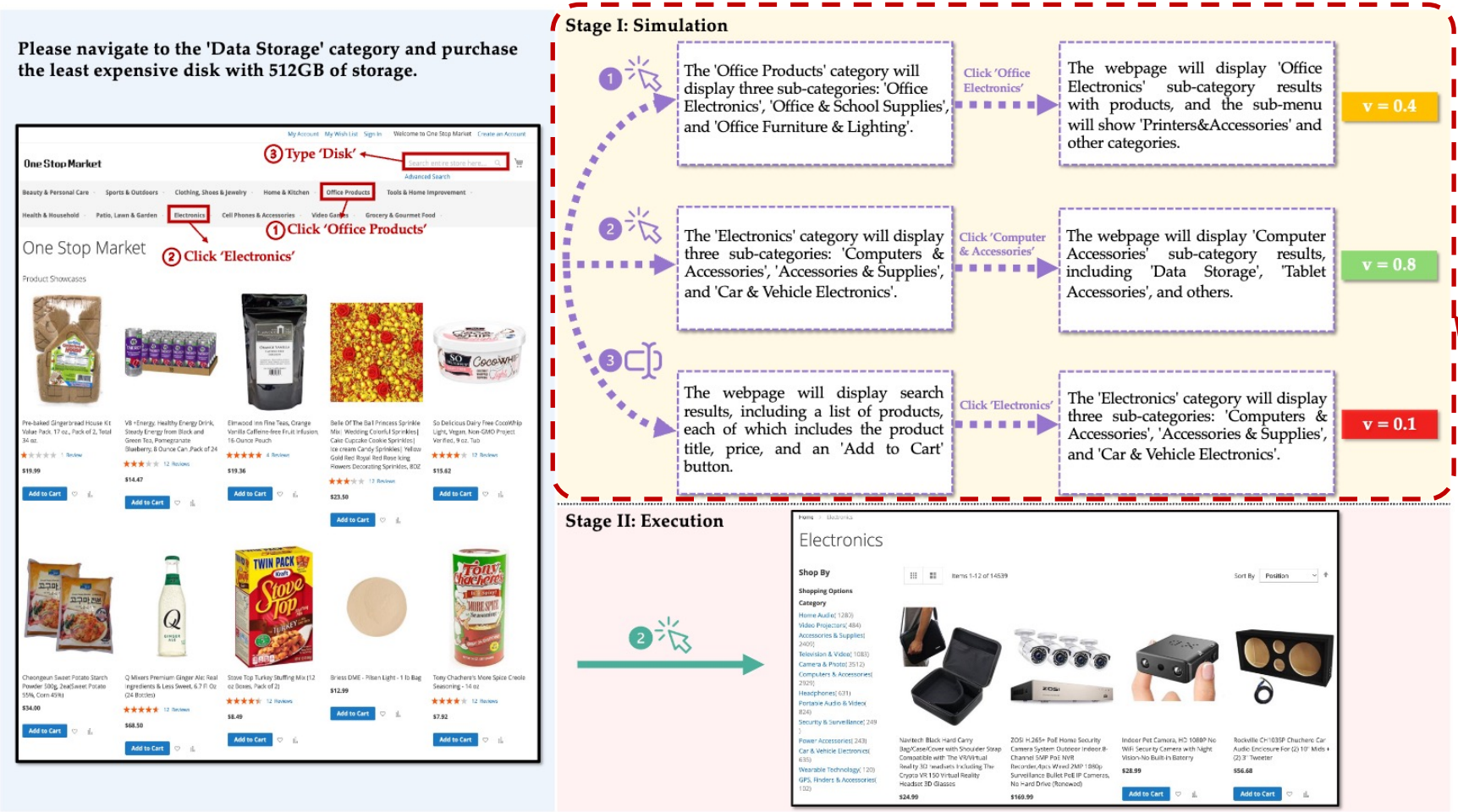


Locally optimal actions ✗ Backtracking is infeasible ✗ Backtracking is infeasible ✓

Gu, Yu, et al. "Is your LLM secretly a world model of the internet? model-based planning for web agents." arXiv preprint arXiv:2411.06559 (2024).

Action Reasoning – Strategic Reasoning

❑ WebDreamer



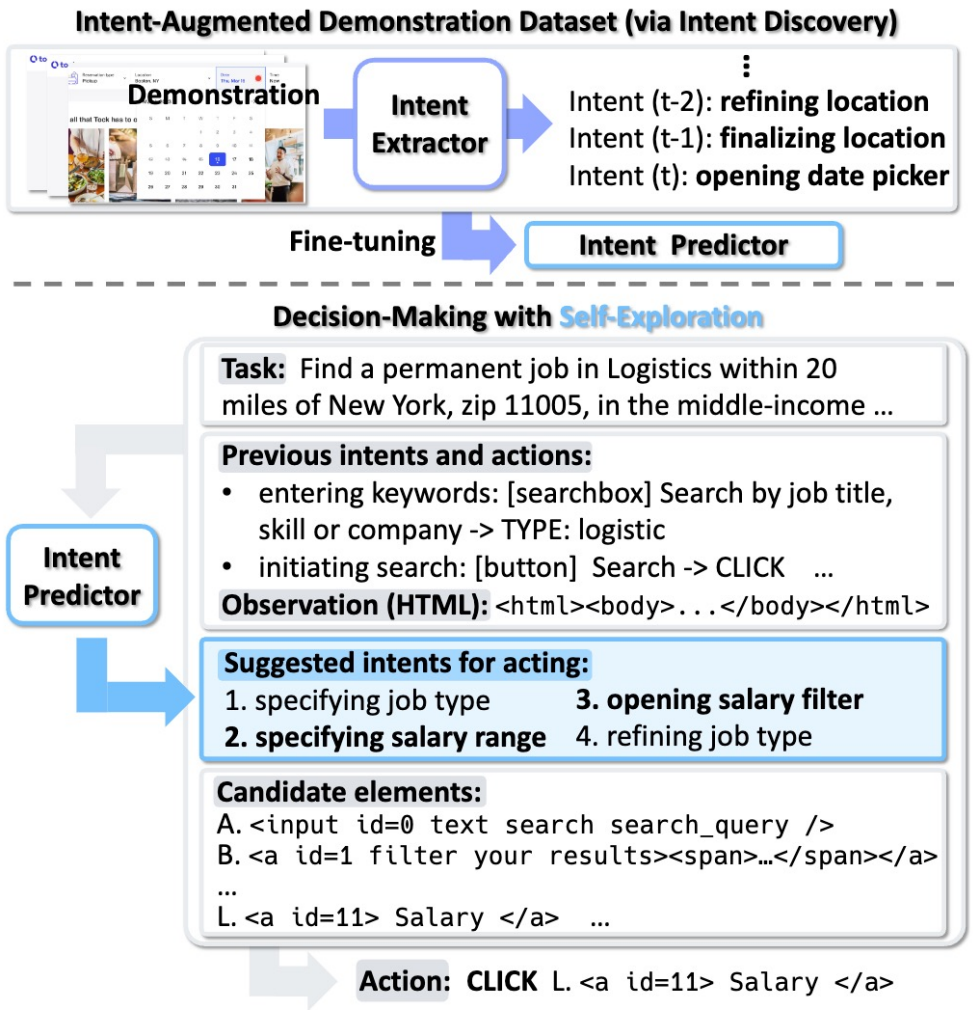
Algorithm 1: WEBDREAMER

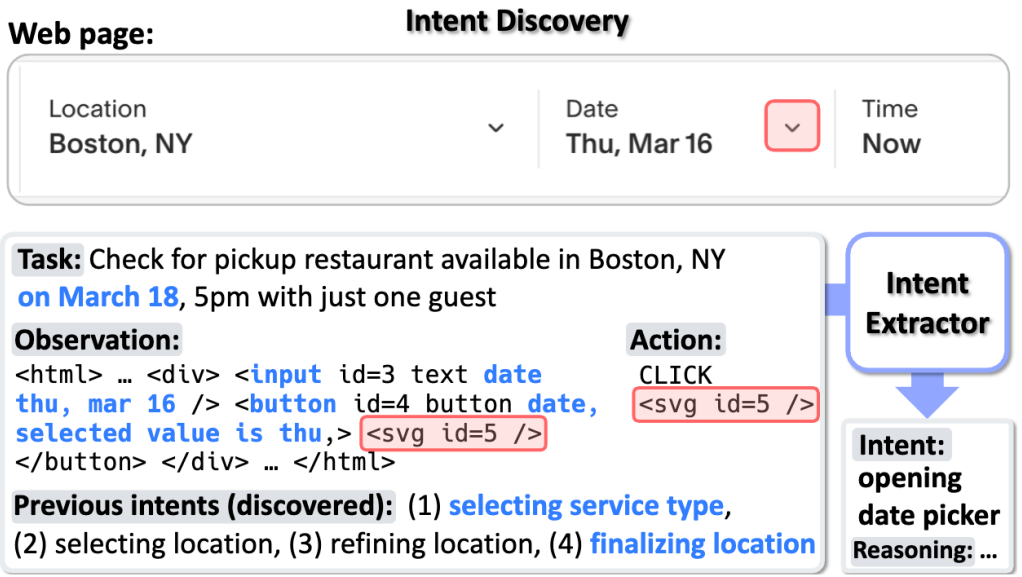
Input: Instruction I; initial observation o_0
Output: Sequence of actions $a_0, a_1, \ldots, a_T$
$t \leftarrow 0$;
while *True* **do**
  $\mathcal{A}_t \leftarrow \texttt{get\_candidate}(I, o_t)$;
  $\mathcal{A}'_t \leftarrow \texttt{self\_refine}(\mathcal{A}_t)$;
  $a_t = \arg\max_{a \in \mathcal{A}'_t} \texttt{score}(\texttt{sim}(o_t, a))$;
  $o_{t+1} \leftarrow \texttt{execute}(a_t)$;
  $t \leftarrow t+1$;
  if `termination_check()` = *True* **then**
    break;
  end
end
Return result;

Gu, Yu, et al. "Is your LLM secretly a world model of the internet? model-based planning for web agents." arXiv preprint arXiv:2411.06559 (2024).

Action Reasoning – Strategic Reasoning

❑ Auto-Intent





Intent Extractor: $z_t = \mathcal{M}_{\text{extract}}(o_t, a_t, z_{1:t-1})$

Intent Prediction:

$$\hat{z}_t^1, \ldots, \hat{z}_t^k \sim \mathcal{M}_{\text{intent}}(o_t, a_{1:t-1}, z_{1:t-1})$$

Action Prediction: $a_t = \pi(o_{1:t}, a_{1:t-1}, \hat{z}_t^{1:k})$.

Kim, Jaekyeom, et al. "Auto-Intent: Automated Intent Discovery and Self-Exploration for Large Language Model Web Agents." EMNLP, 2024.

Reasoning

❑ Reactive Planning

- Efficient
- Easy to implement
- …

- Performance is highly dependent on the capabilities of LFMs
- …

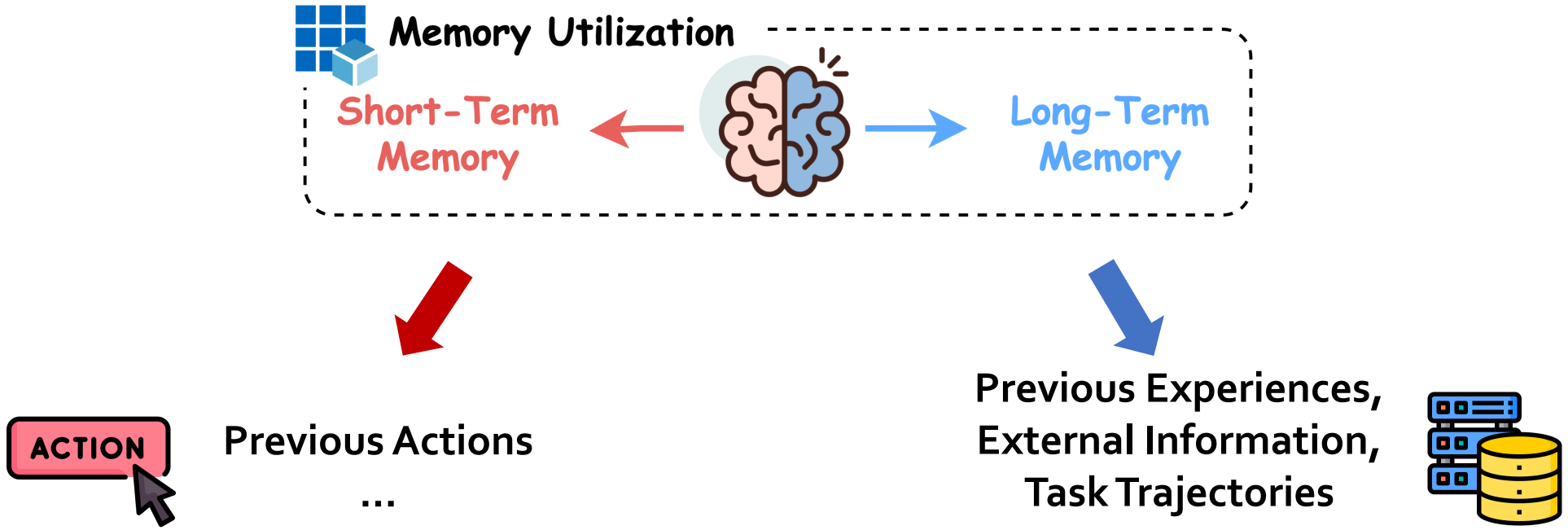
❑ Strategic Reasoning

- More accurate action prediction
- …

- Additional computational cost
- Design sophisticated mechanisms.
- …

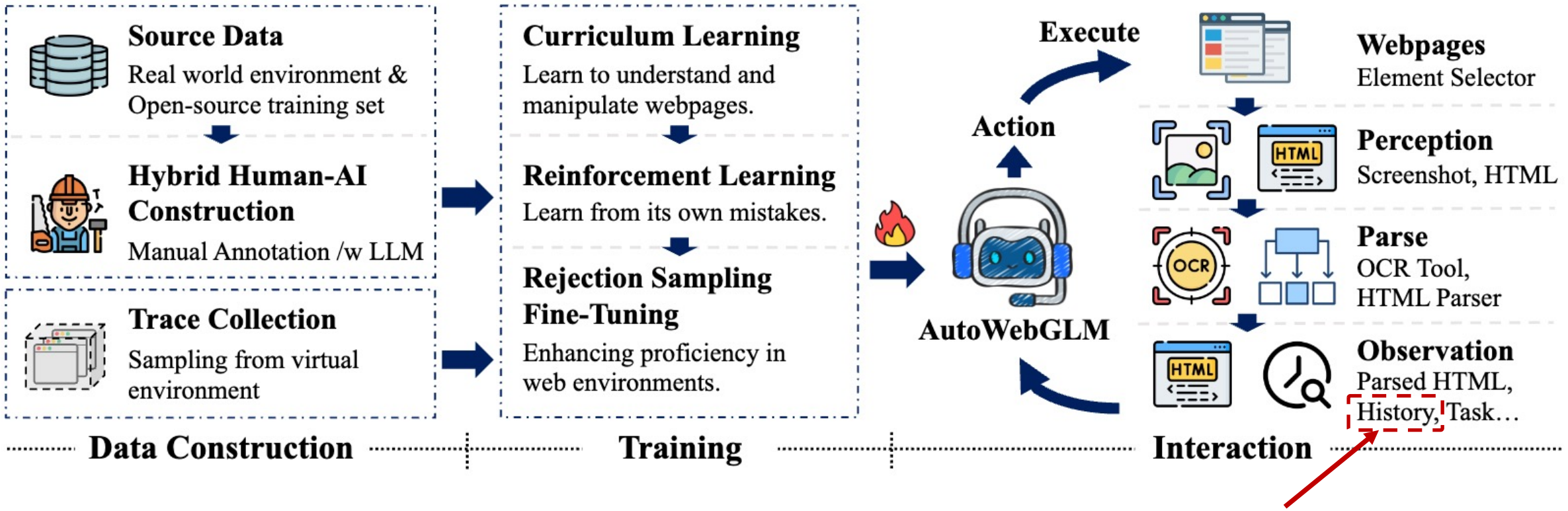
Memory Utilization

- 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



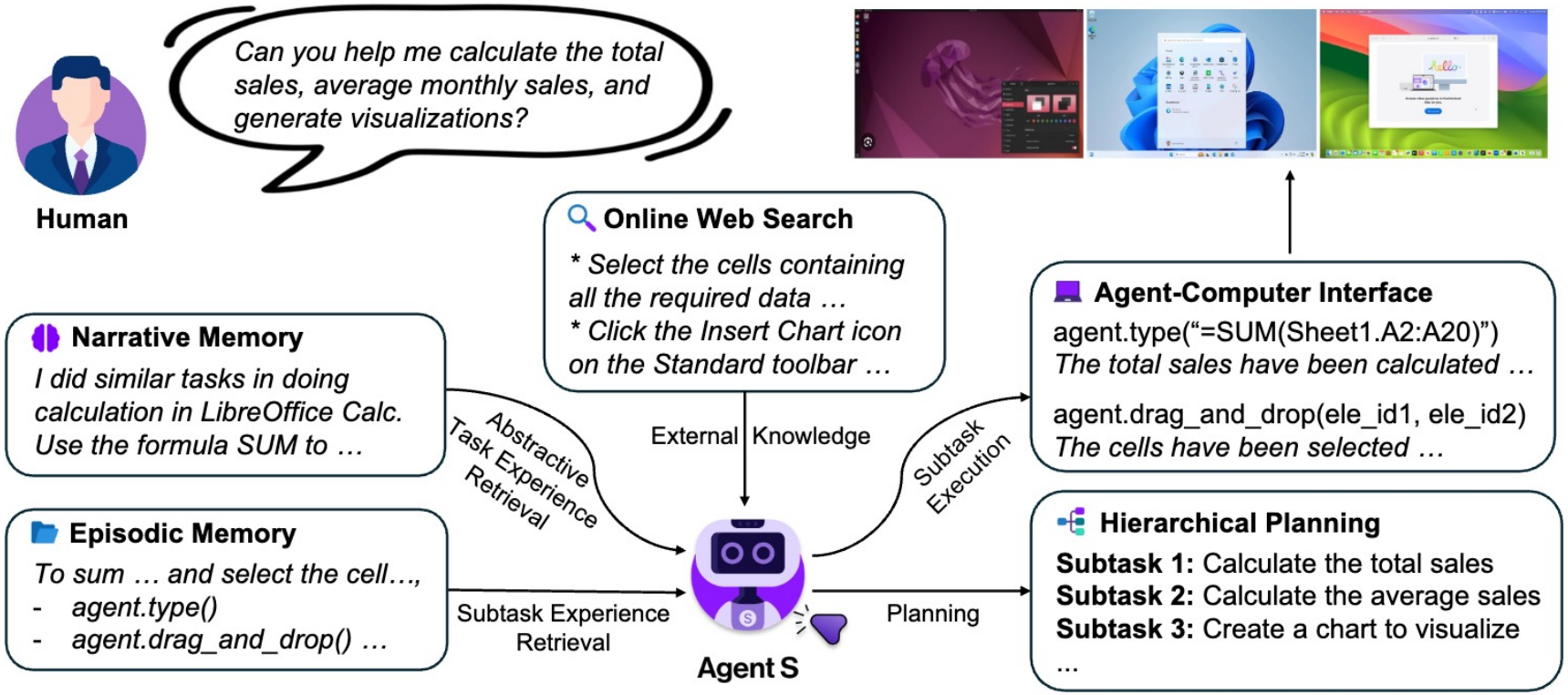
The best solution to inform the agent of past operations is explicitly providing it.

Lai, Hanyu, et al. "AutoWebGLM: A Large Language Model-based Web Navigating Agent." KDD, 2024.

Memory Utilization – Long-Term

❑ AGENT S

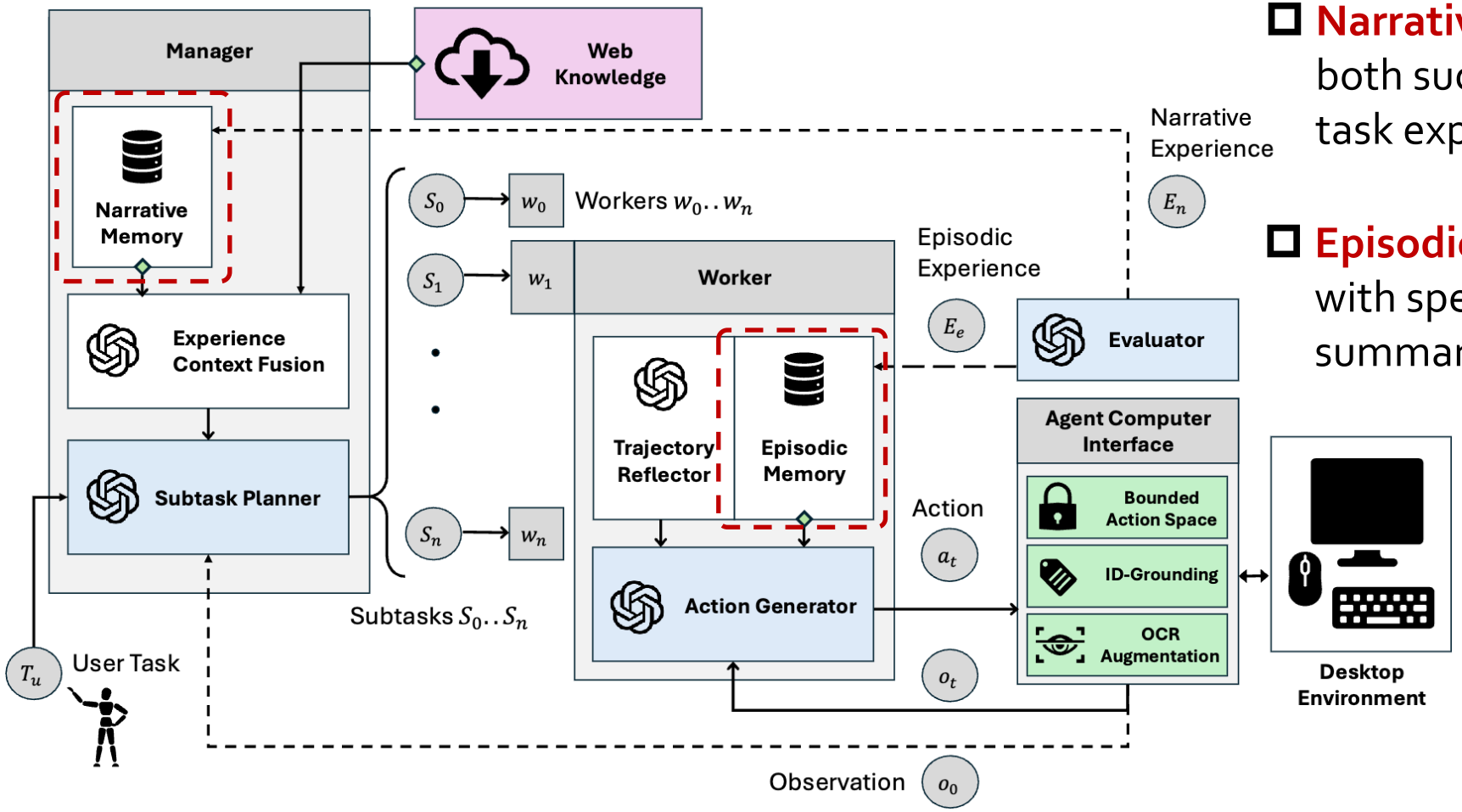
- Leverages **Online Web Knowledge** and past experiences stored in **Narrative Memory** to decompose the complex, long-horizon task into a structured plan of manageable subtasks
- Retrieves step-by-step subtask experience from **Episodic Memory** to refine the actions.



Agashe, Saaket, et al. "Agent S: An Open Agentic Framework that Uses Computers Like a Human." ICLR, 2024.

Memory Utilization – Long-Term

❑ AGENT S

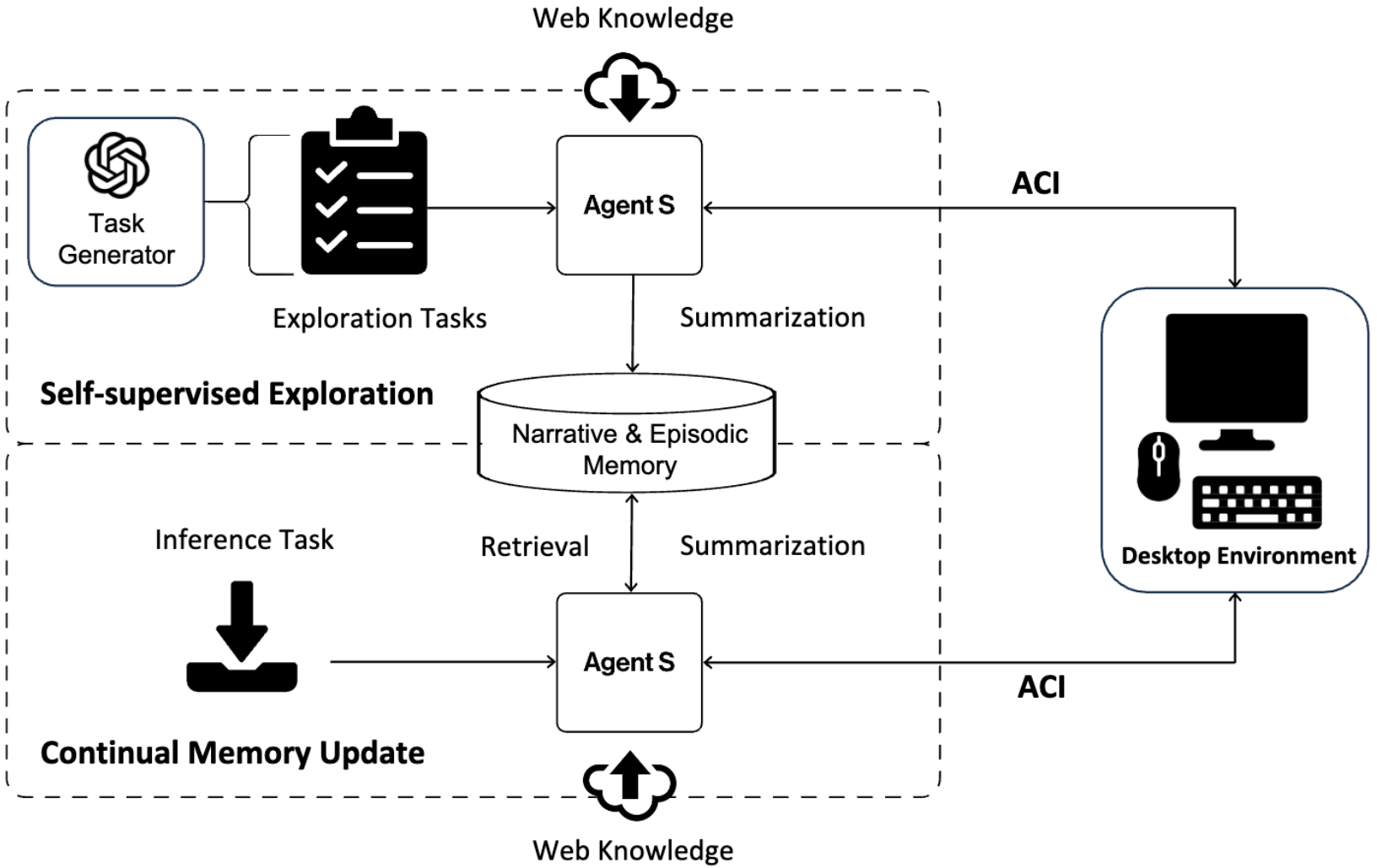


- **Narrative Memory:** includes summaries of both successful and failed trajectories as task experiences.
- **Episodic Memory:** includes a complete plan with specific grounding actions and only summaries from the subtask trajectories

Agashe, Saaket, et al. "Agent S: An Open Agentic Framework that Uses Computers Like a Human." ICLR, 2024.

Memory Utilization – Long-Term

❑ AGENT S



Agashe, Saaket, et al. “Agent S: An Open Agentic Framework that Uses Computers Like a Human.” ICLR, 2024.

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
- ➢ …

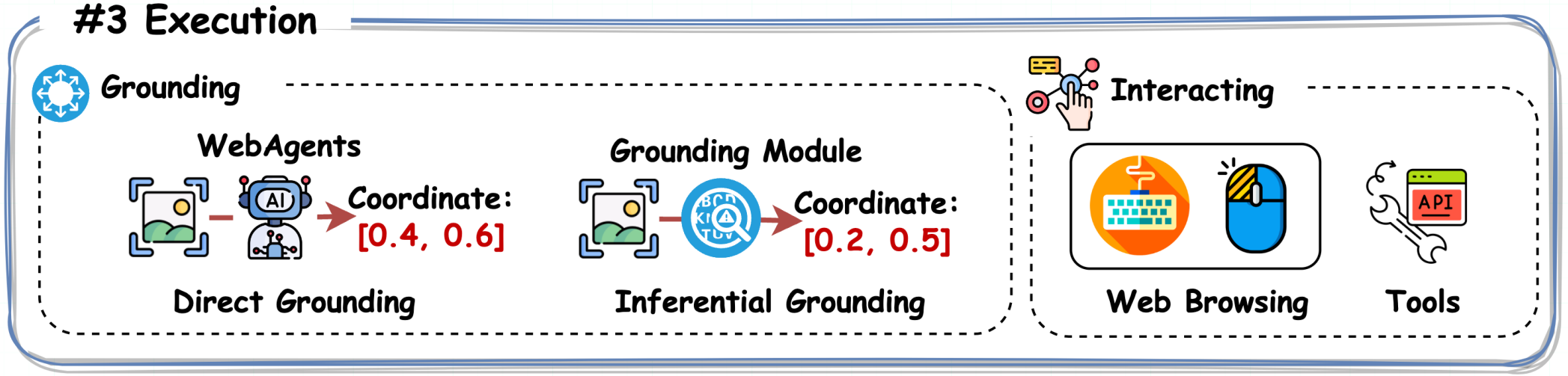
PART 3: Architectures of WebAgents

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**

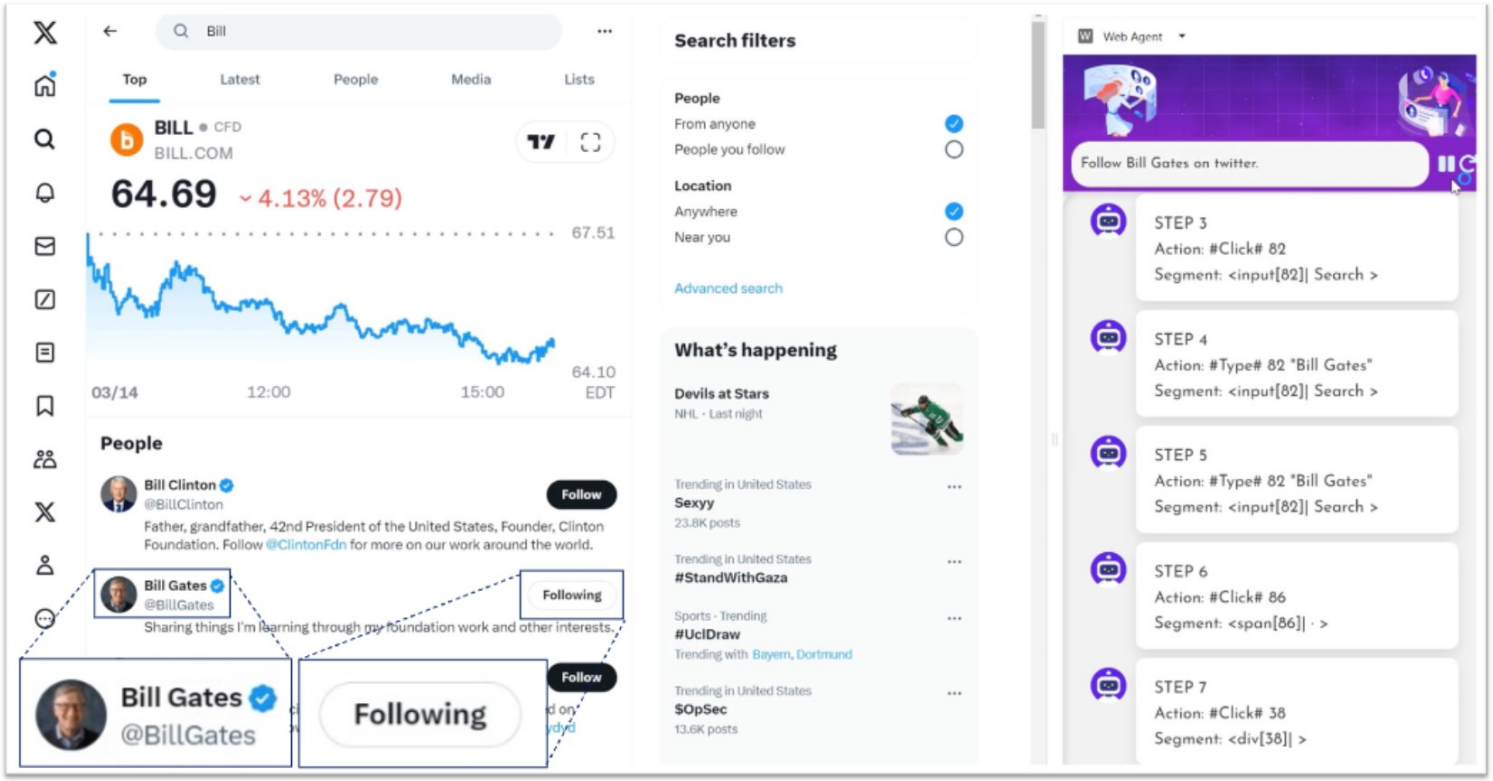
Execution

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

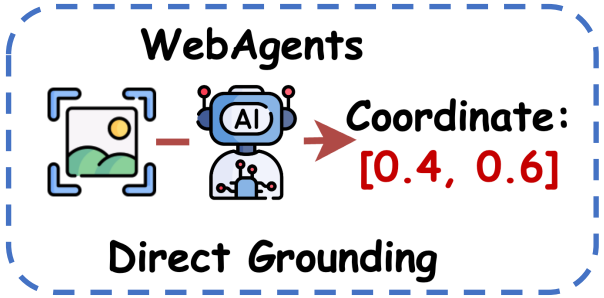


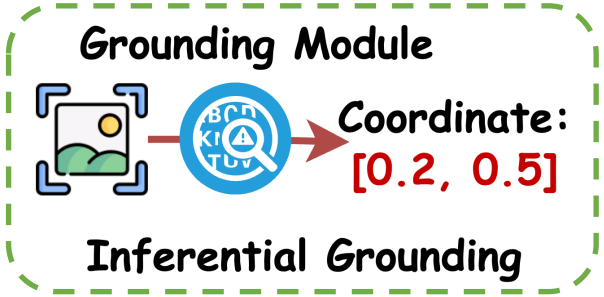
Grounding

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



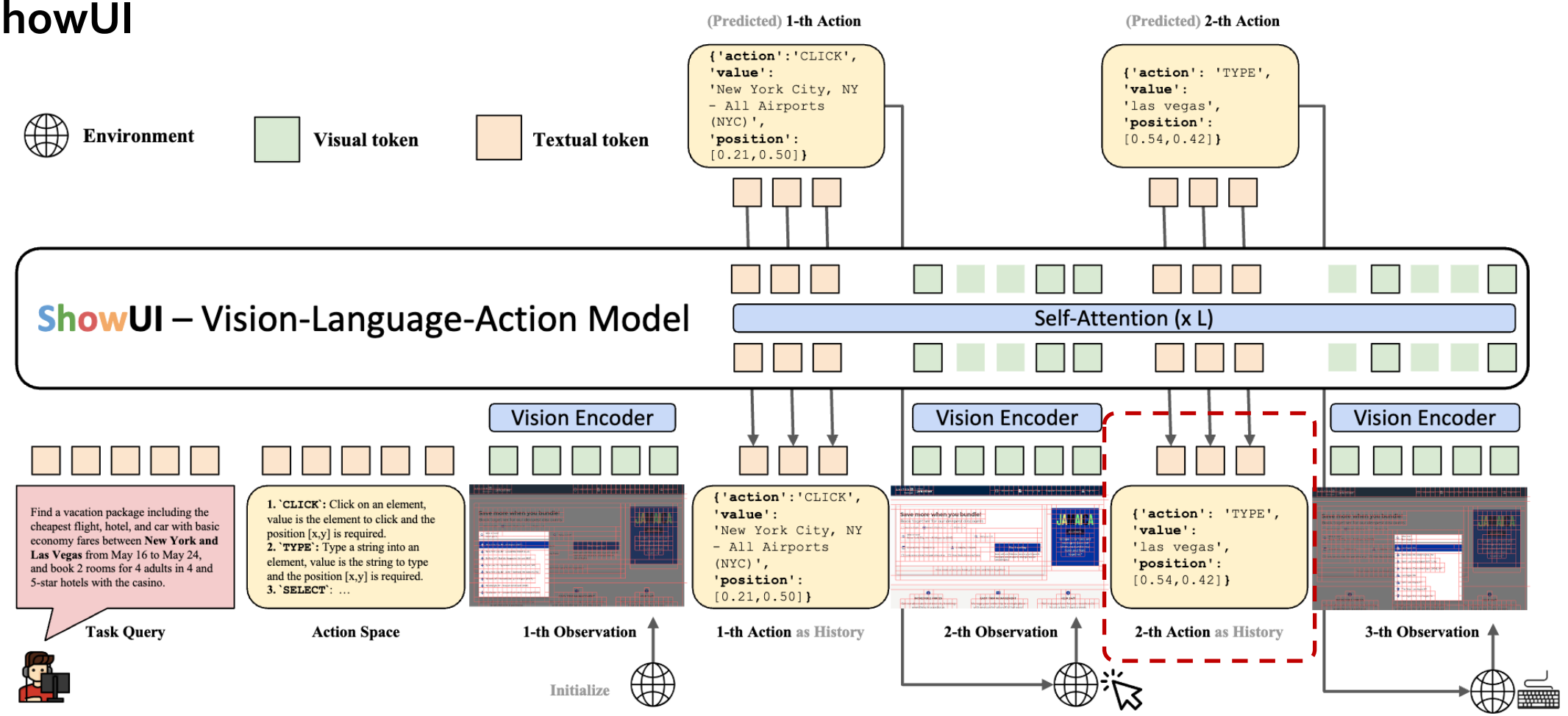
(a) Follow Bill Gates on X (Twitter).





Grounding – Direct Grounding

❑ ShowUI



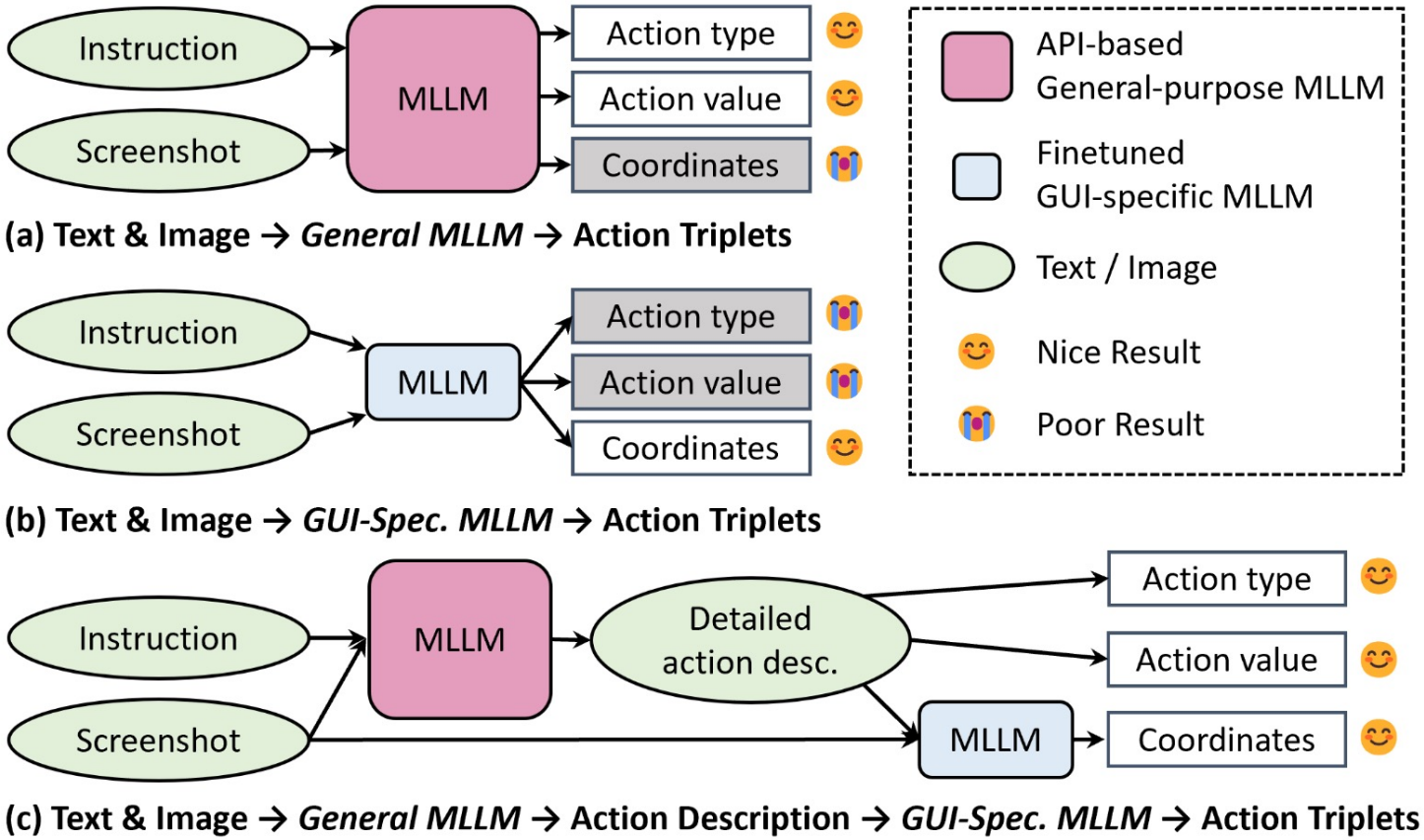
Contains both generated actions and the target element position

Lin, Kevin Qinghong, et al. "Showui: One vision-language-action model for GUI visual agent." CVPR, 2025.

Grounding – Inferential Grounding

❑ Ponder & Press

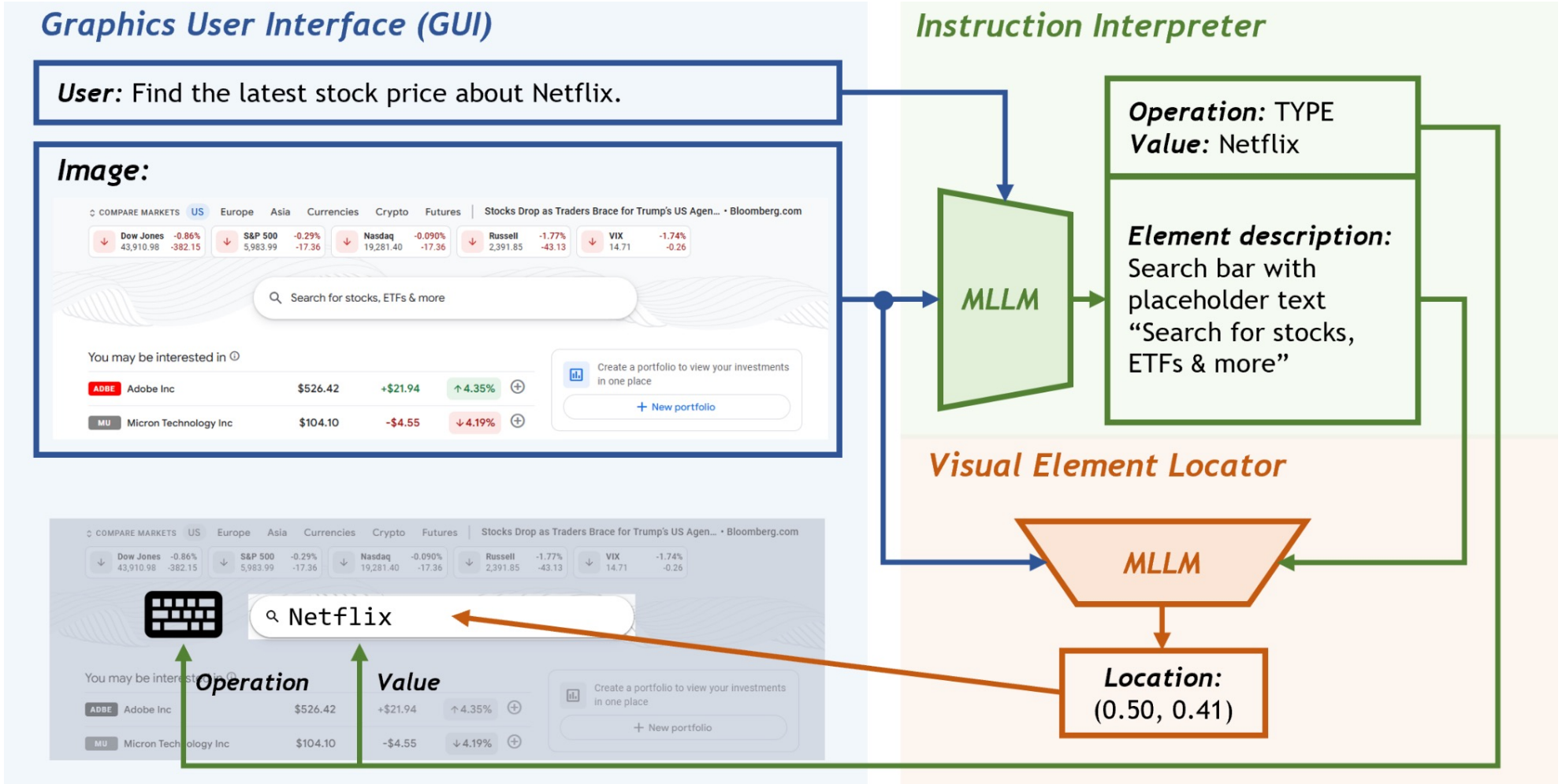
User instructions are directly mapped to action triplets in a single inference, which is difficult due to the significant difference between the textual nature of actions and values, and the numerical nature of pixel coordinates.



Wang, Yiqin, et al. "Ponder & press: Advancing visual gui agent towards general computer control." arXiv preprint arXiv:2412.01268 (2024).

Grounding – Inferential Grounding

❑ Ponder & Press



Wang, Yiqin, et al. "Ponder & press: Advancing visual gui agent towards general computer control." arXiv preprint arXiv:2412.01268 (2024).

Grounding

❑ Direct Grounding

- ➢ Efficient
- ➢ Easy to implement
- ➢ …

- ➢ Grounding accuracy is dependent on the capabilities of LFMs
- ➢ …

❑ Inferential Grounding

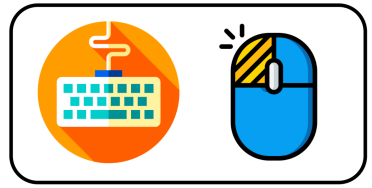
- ➢ More accurate element location
- ➢ …

- ➢ 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.**



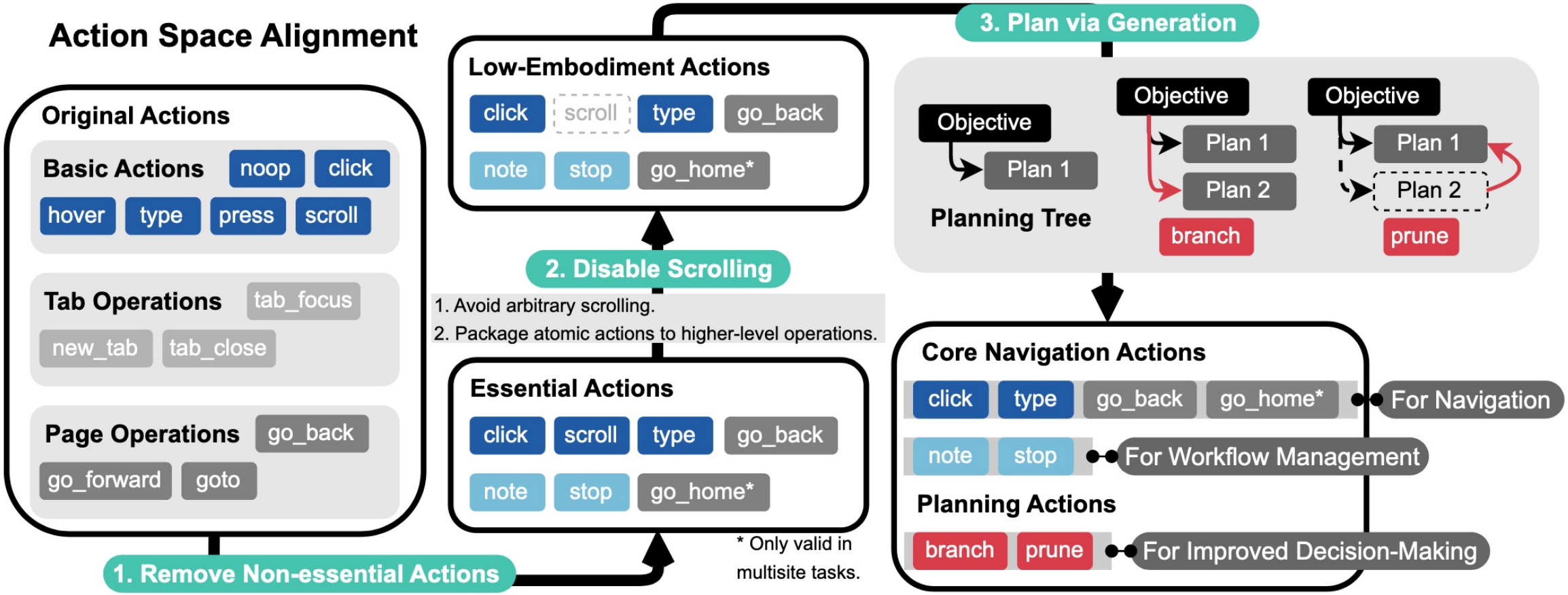
Web Browsing

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

Tools

Interacting – Web Browsing

❑ AGENTOCCAM

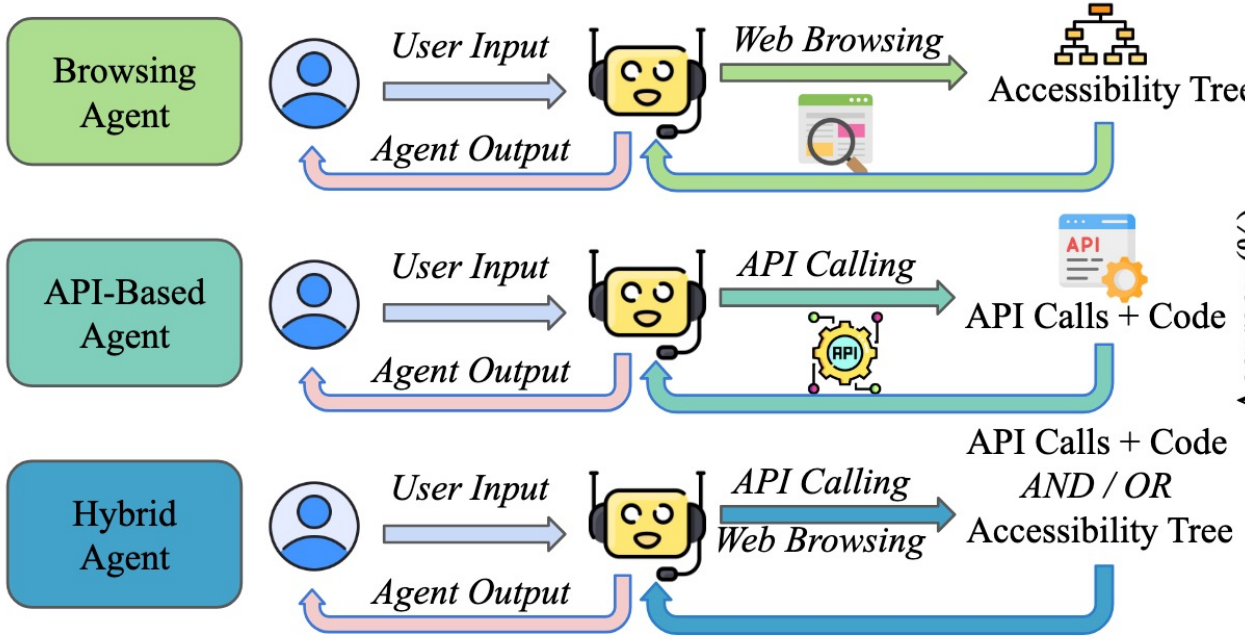


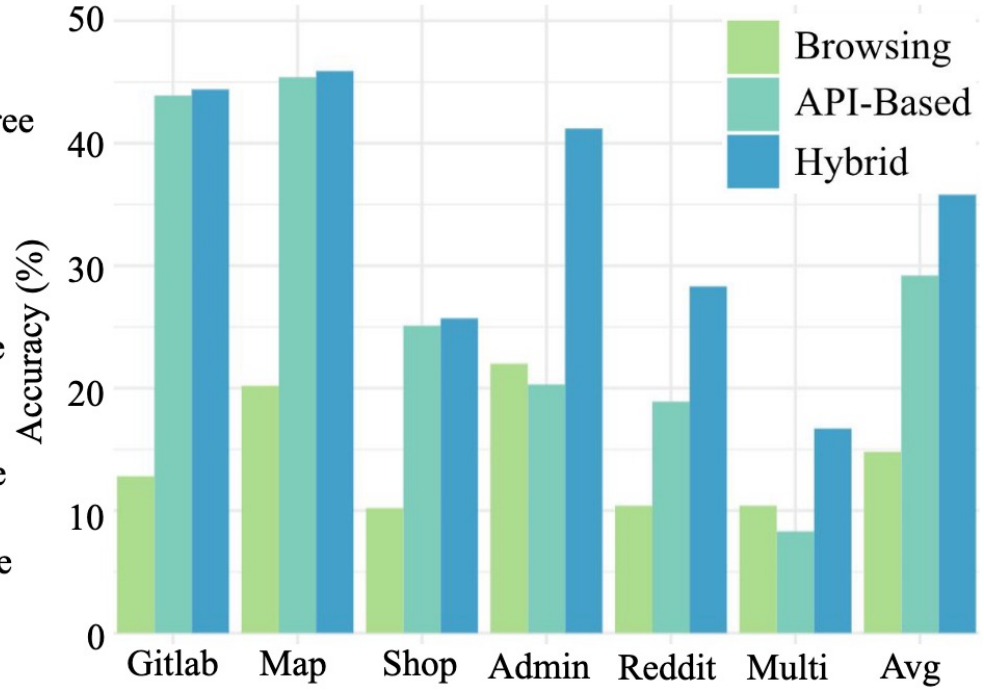
Yang, Ke, et al. "AgentOccam: A Simple Yet Strong Baseline for LLM-Based Web Agents." ICLR, 2024.

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.**





Song, Yueqi, et al. "Beyond browsing: API-based web agents." arXiv preprint arXiv:2410.16464 (2024).

Interacting – Tool

❑ API-Calling Agents


```
(1) goto `gitlab.com` (2) login with credentials
(3) click `a11yproject` (4) click `Repository`
(5) click `Commits` (6) No commits found -> scroll down
(7) No commits found -> scroll down ..... (15) No
commits found but no steps left, conclude 0 commits are
made by SaptakS.
```

Web Browsing Traces. Failed after 15 steps.

```
<execute_ipython>
(1)r=requests.get('gitlab.com/api/a11yproject/commits')
(2)commits=r.json()
(3)len([c for c in commits if c['author'] == 'SaptakS'])
</execute_ipython>
```

API Calling via Python requests library

API Documentation

```
# 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.
```

API Calling

```
<execute_ipython>
requests.get('gitlab.com/api/a11yproject/commits')
</execute_ipython>
```

JSON Output

```
[ ......{
    "id": "ed37a2f2",
    "created_at": "2023-03-13T21:04:49.000-04:00",
    "title": "Update README.md",
    "message": "Update README.md",
    "author": "SaptakS",
}]
```


Song, Yueqi, et al. "Beyond browsing: API-based web agents." arXiv preprint arXiv:2410.16464 (2024).

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
- ➢ …