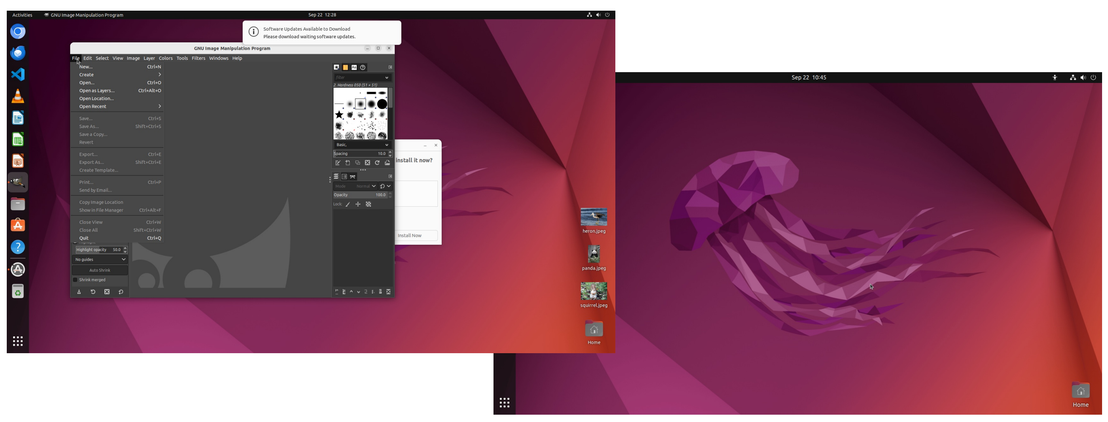
Agent S : Retrieval-as-Learning for GUI agents

Saaket Agashe*, Jiuzhou Han*, Shuyu Gan, Jiachen Yang, Ang Li, Xin Eric Wang

Introduction



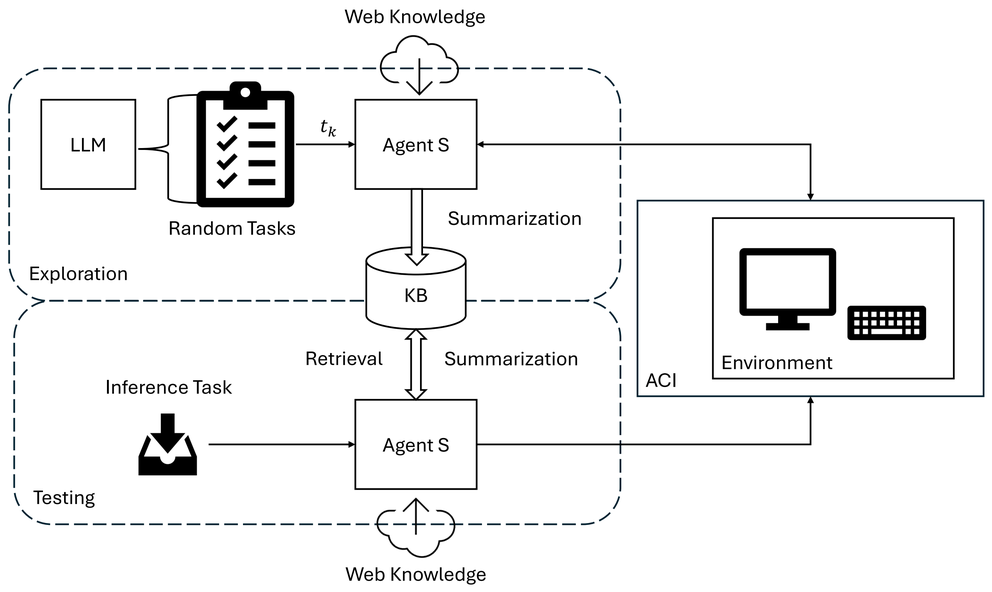
- Can autonomous agents take control of your keyboard and mouse and help you automate those mundane tasks?

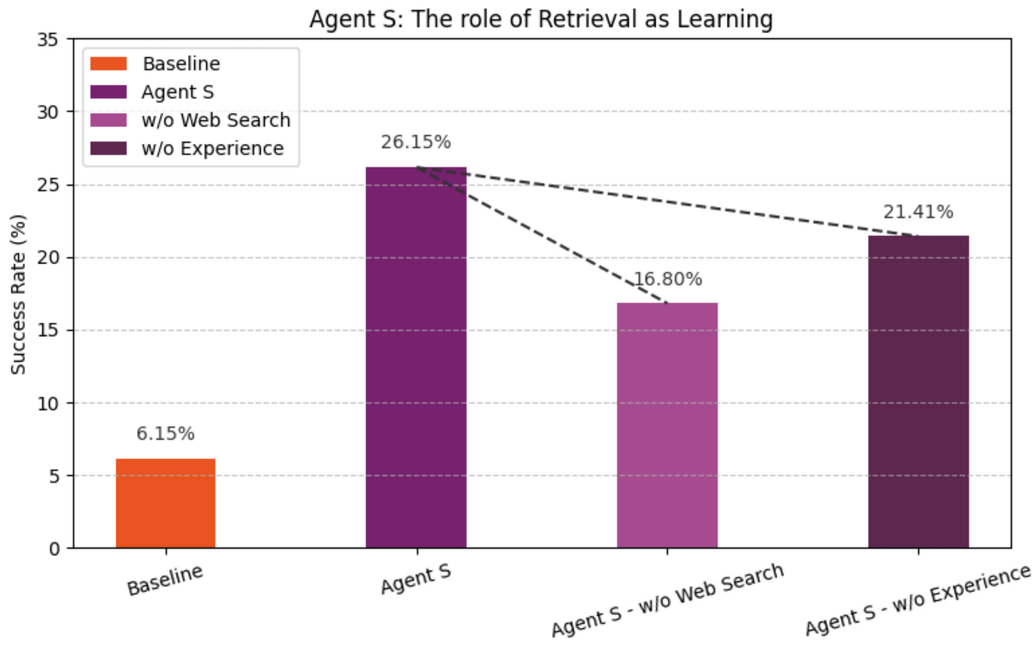
- Recent advancements in Large Multimodal Models (LMMs) have reignited interest in developing fully autonomous agents capable of operating in human-centered interactive systems, particularly desktop OS environments.

- However, developing such agents remains extremely challenging due to:

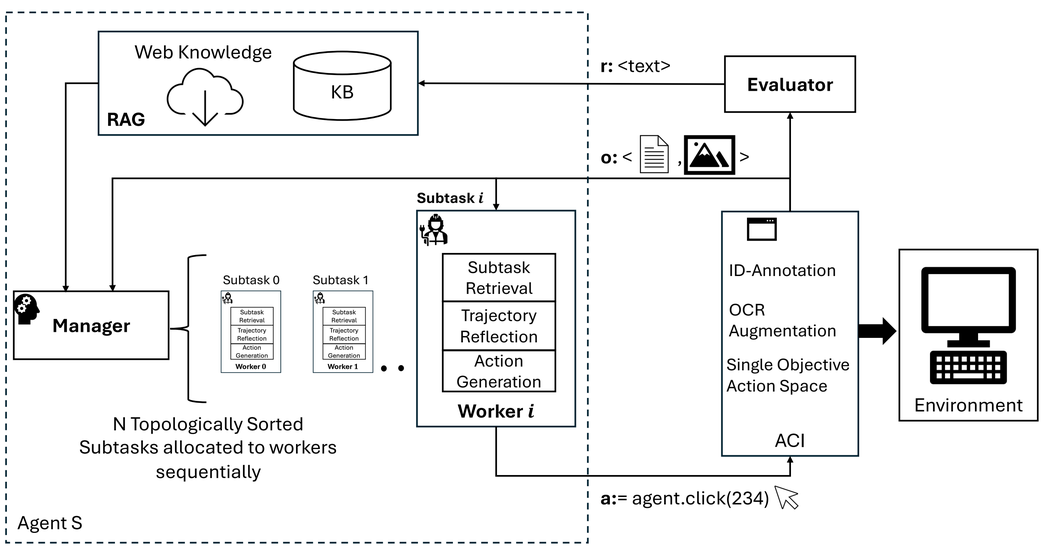
  1. The vast diversity of screens, apps, UI elements, and interactions that exist across different platforms and software.
  2. The dynamic nature of GUIs, including frequent updates and real-time changes in interface layouts.
  3. The difficulty in adapting general-purpose LMMs to act as specialized GUI agents, given the lack of domain-specific knowledge and training data.

Experience Retrieval as Learning

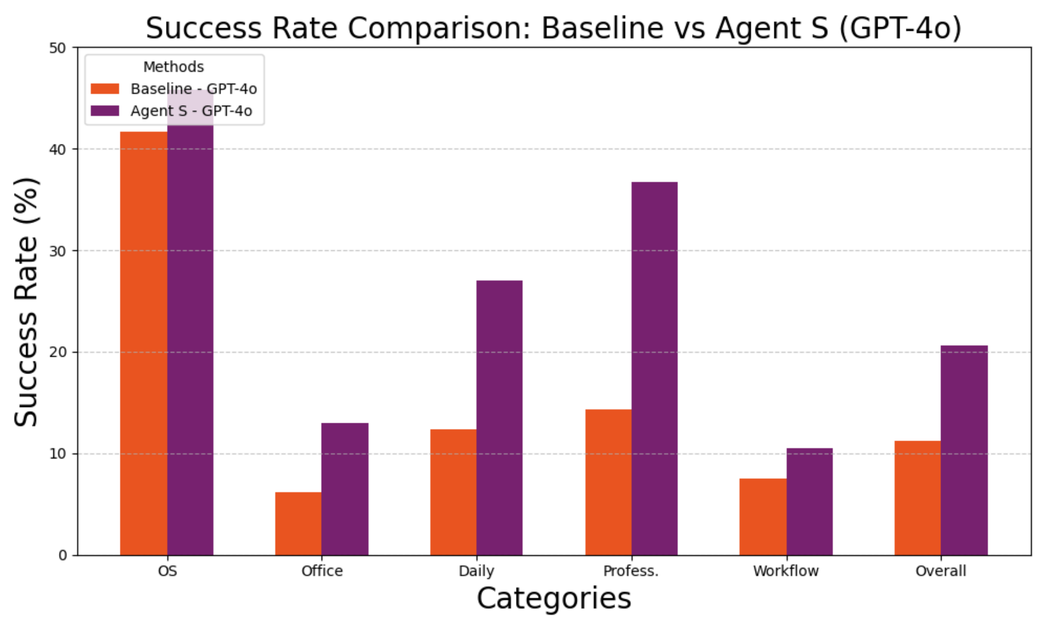




The Agent S Framework



Results in OSWorld



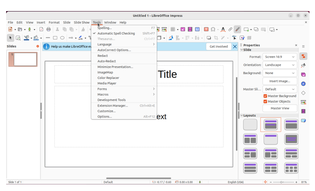
- **Hierarchical Planning and Learning:** Agent S uses a manager for high-level planning and workers for executing subtasks in GUI problems.

- **Agent-Computer Interface:** Agent S interacts with an abstraction layer, simplifying learning through annotated input, focused actions, and rich feedback.

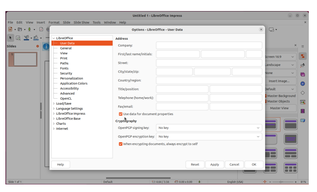
- **Web Knowledge:** Agent S utilizes online resources to gain domain knowledge for GUI tasks.

- **Retrieval as Learning:** Agent S learns by storing experiences and retrieving relevant information as needed.

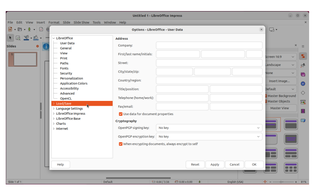
Agent S in Action



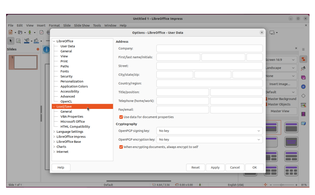
(a) Click on the *Tools* menu: agent.click(38, 1, left)



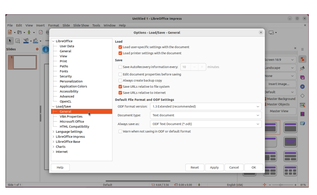
(b) Click on the *Options...* item: agent.click(53, 1, left)



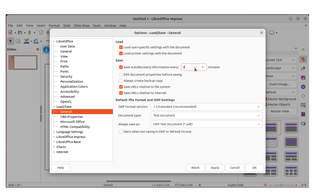
(c) Click on *Load/Save* category: agent.click(207, 1, left)



(d) Double-click on *Load/Save* category: agent.click(207, 2, left)



(e) Click on the General sub-option: agent.click(208, 1, left)



(f) Change the time to 3 minutes: agent.type(230, "3", overwrite=True)