Development of LLM-driven GUI Agents

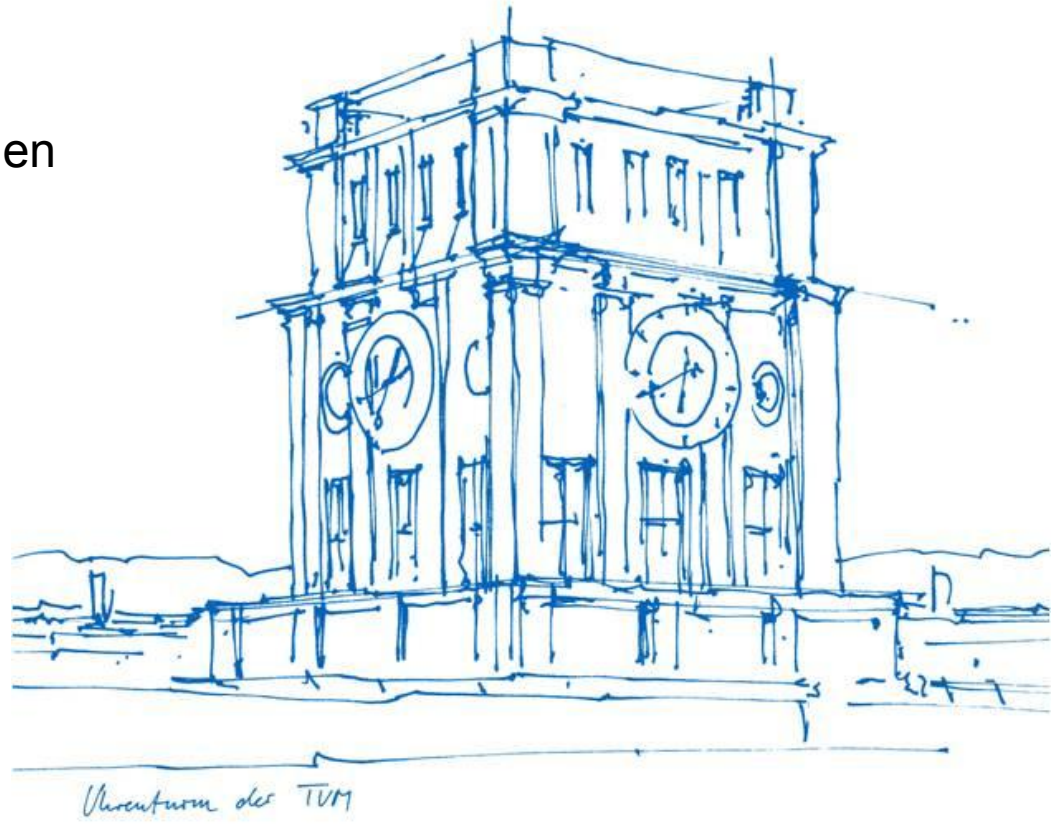
OSWorld

De Lamo Castrillo, Victor

Oliver Suriñach, Santiago

Technische Universität München

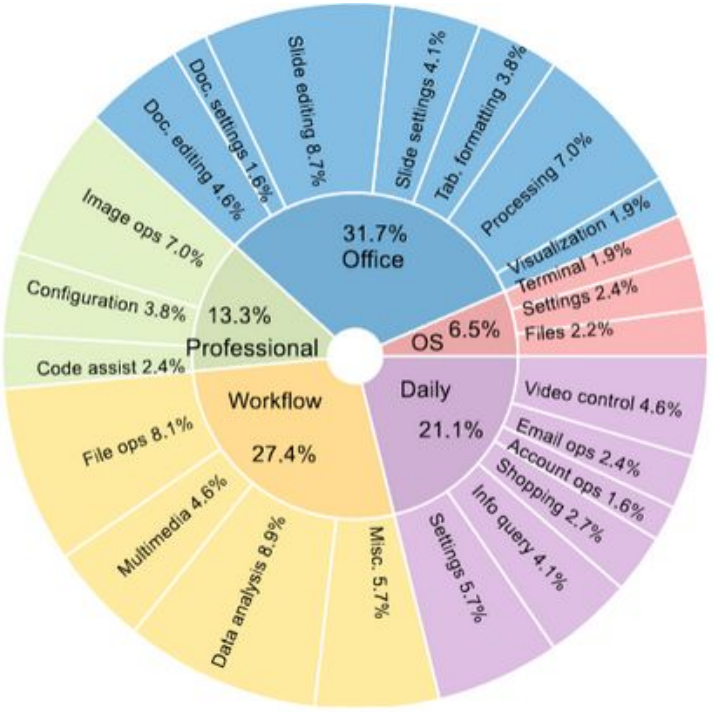
Garching, 2025-07-21



OSWorld Benchmark

369 tasks and 43 tasks on Windows.

- OS
- Daily
- Professional
- Office
- Workflow



Base Agent

Perception Expert

Challenges:

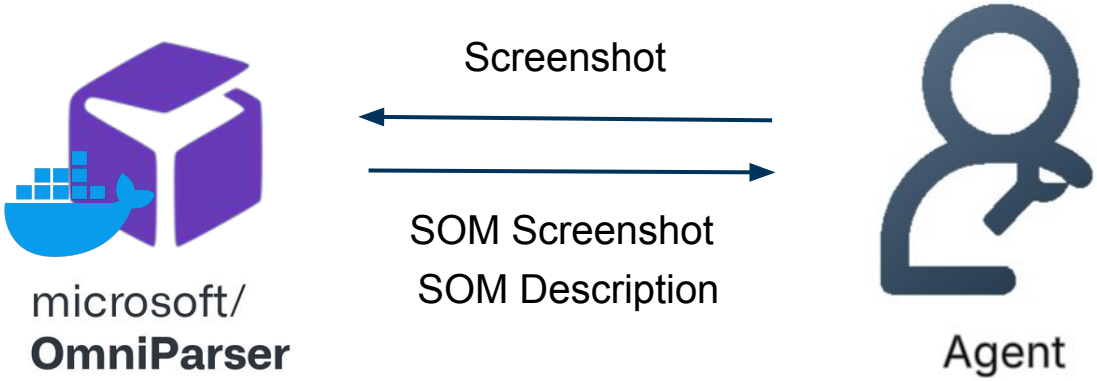
- VLM is not able to get the right coordinates
- The only available object is the vanilla screenshot
- Prompt engineering is not enough

Solution:

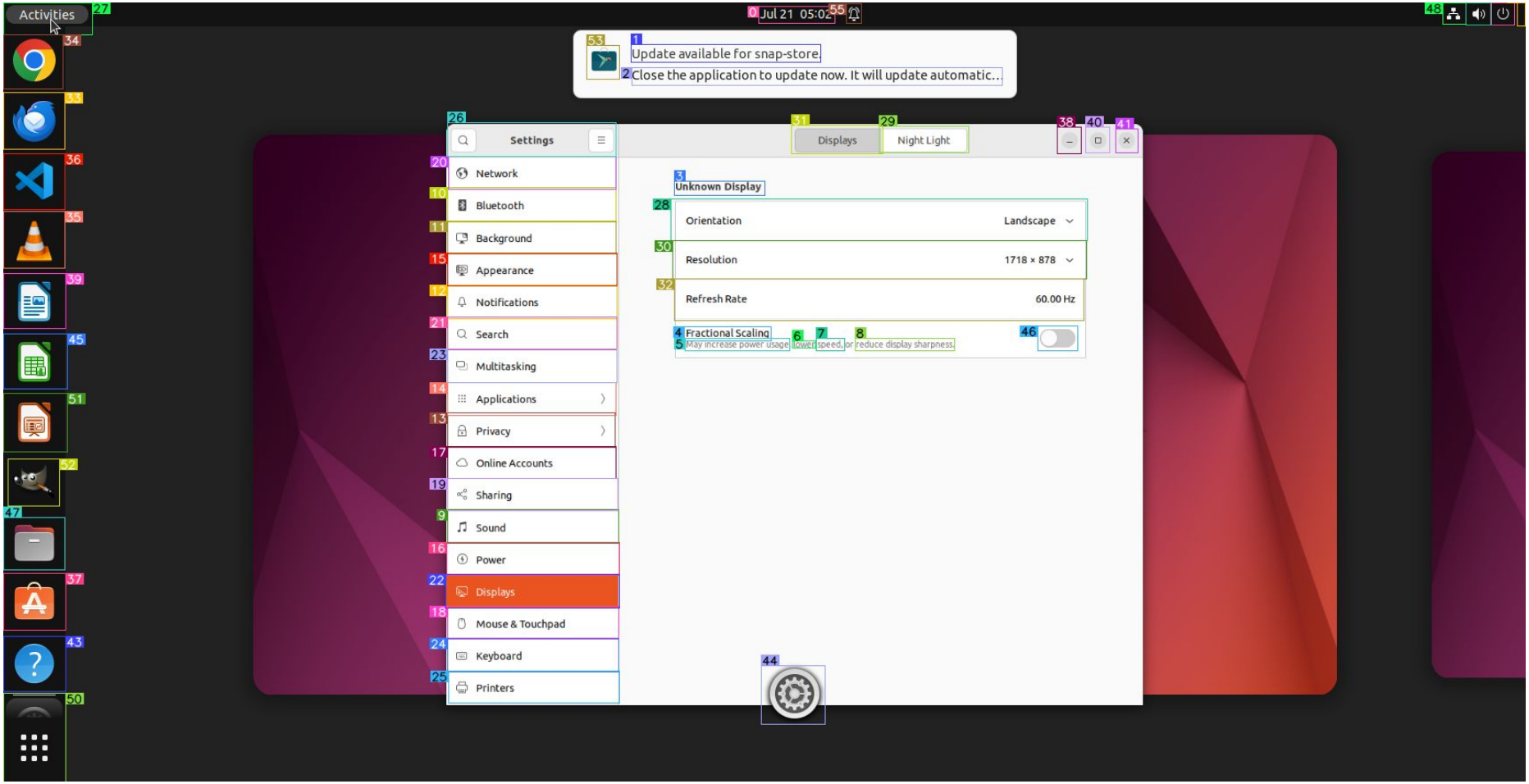
Obtain a better visual object

Perception Expert

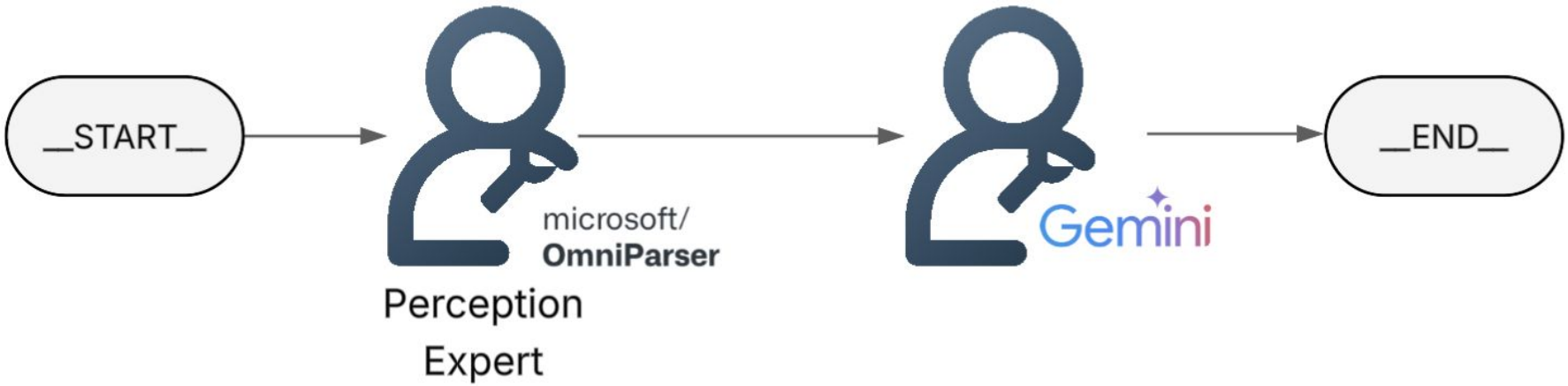
- **Omniparser v2**
  - Fine tuned YOLO and Florence-2
  - Omniserver API



Perception Expert. SOM Screenshot



Perception Expert



Action Expert

Challenges:

- Specific pyautogui instructions needed
- Not enough element recognition
- Lack of action precision

Solution:

Create an action expert, prominent in:

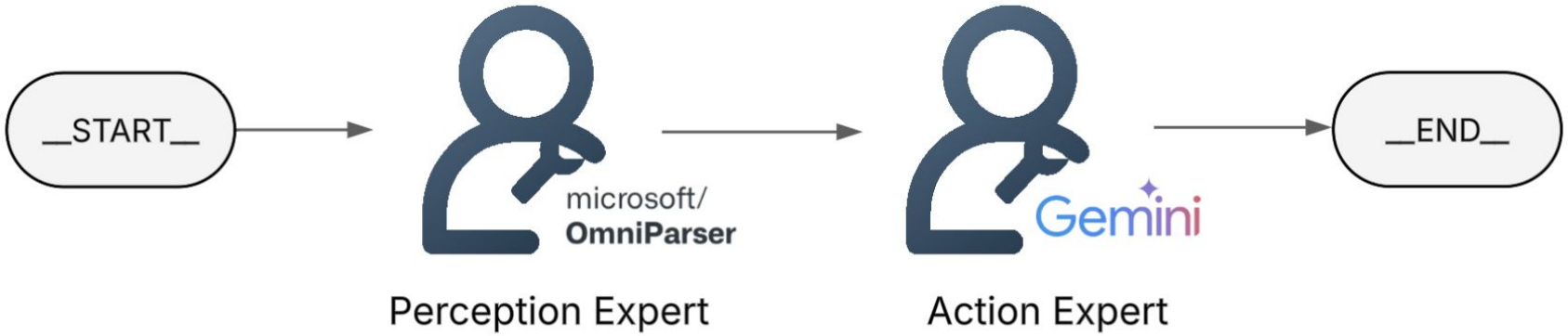
- SOM understanding
- Coordinate accuracy
- Pyautogui code generation

Action Expert

Implementation:

- **COT** prompt methodology
  1. Reflection about solution
  2. Box subset with SOM description
  3. Main box selection
  4. Coordinates transformation
  5. Pyautogui generation

Action Expert



Planning Expert

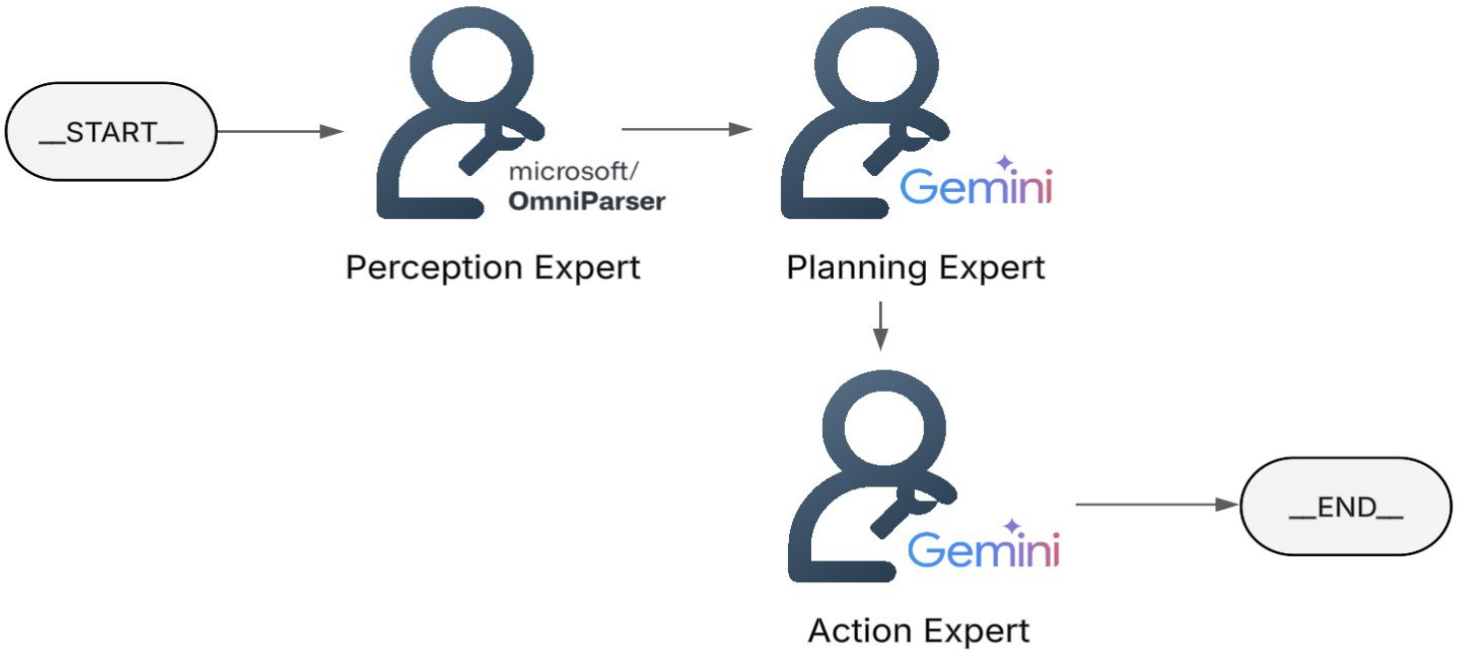
Challenges

- It forgets the main goal
- Not the most optimal plan
- Need of different options to perform the task

Implementation

- Breaks the main task into the first subtask
- Generates a instruction list for each subtask
- Thinks the next subtask when completing one

Planning Expert



Reflection Expert

Challenges

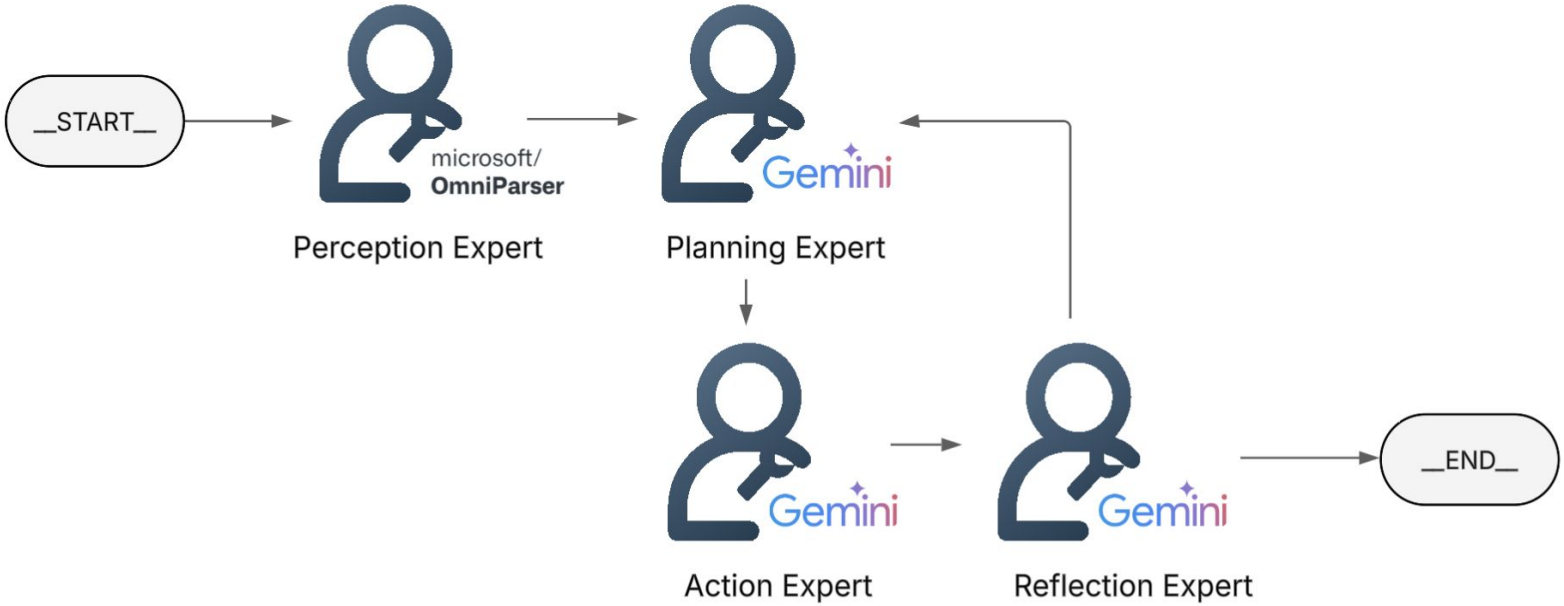
- Evaluation of the correctness of the action
- Always needs the planning expert
- Provide accurate feedback to planning and action

Reflection Expert

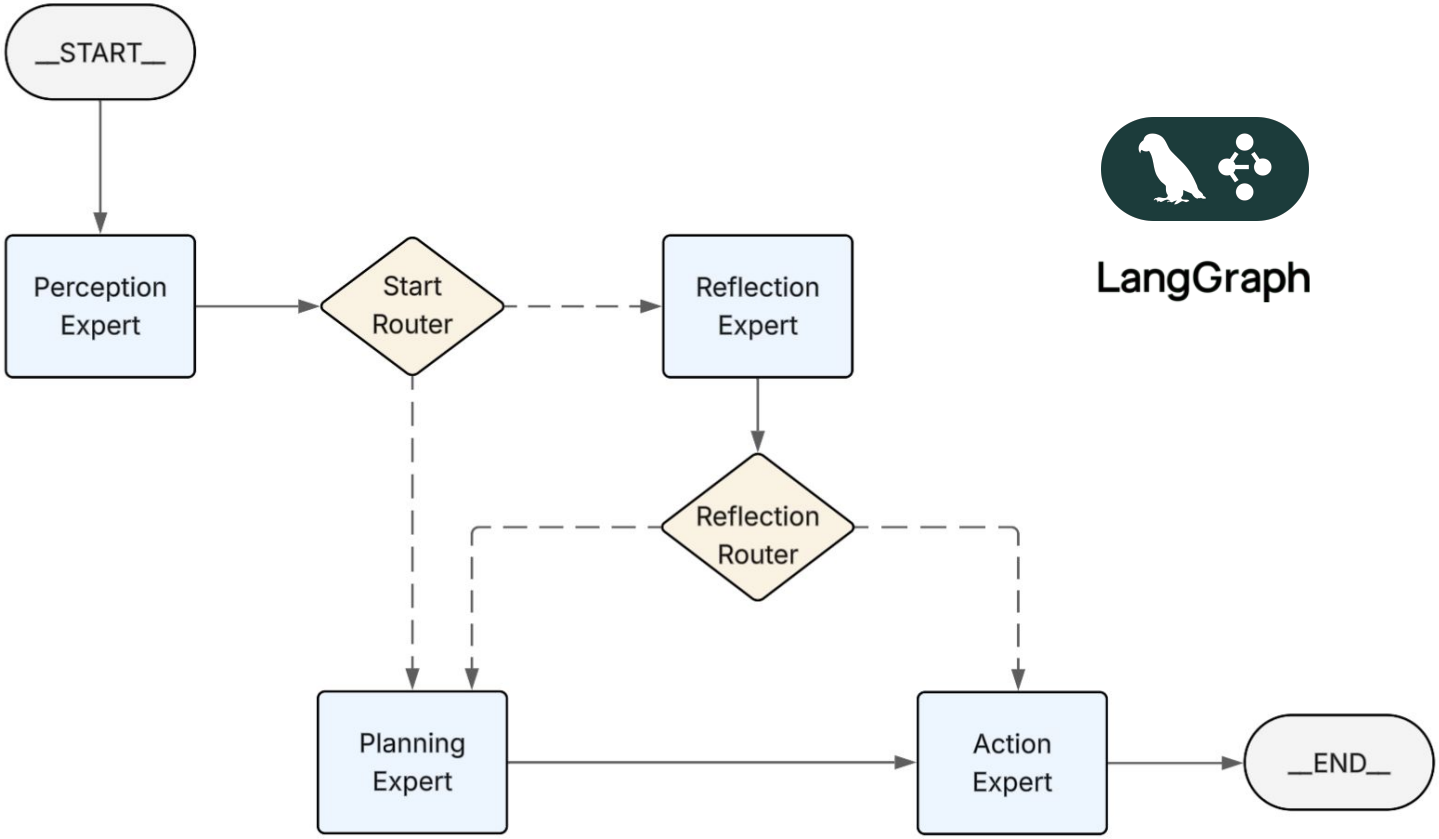
Implementation

- Evaluates the performance of the action expert
- If the action expert is accurate
  - Schedules the next instruction
  - No instructions left, then calls the planning
- If the action expert has failed
  - Classifies the error in minor or major

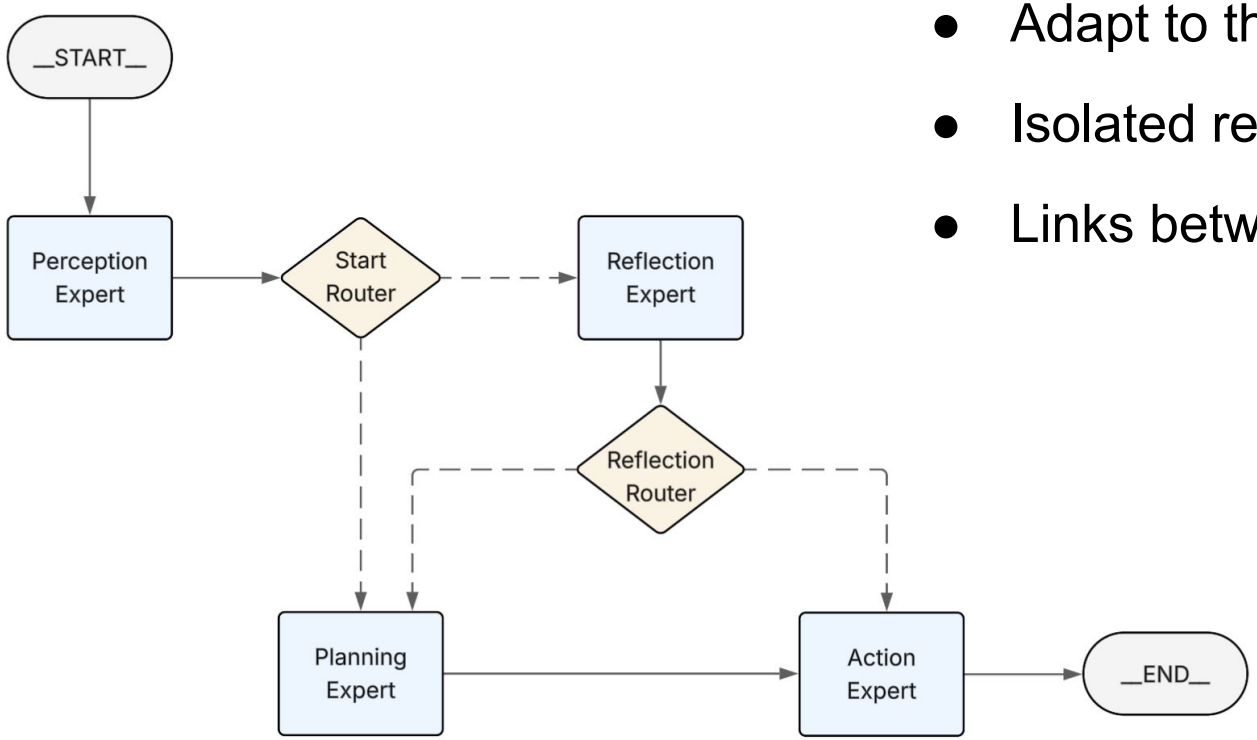
Reflection Expert



Agent Architecture



Agent Architecture



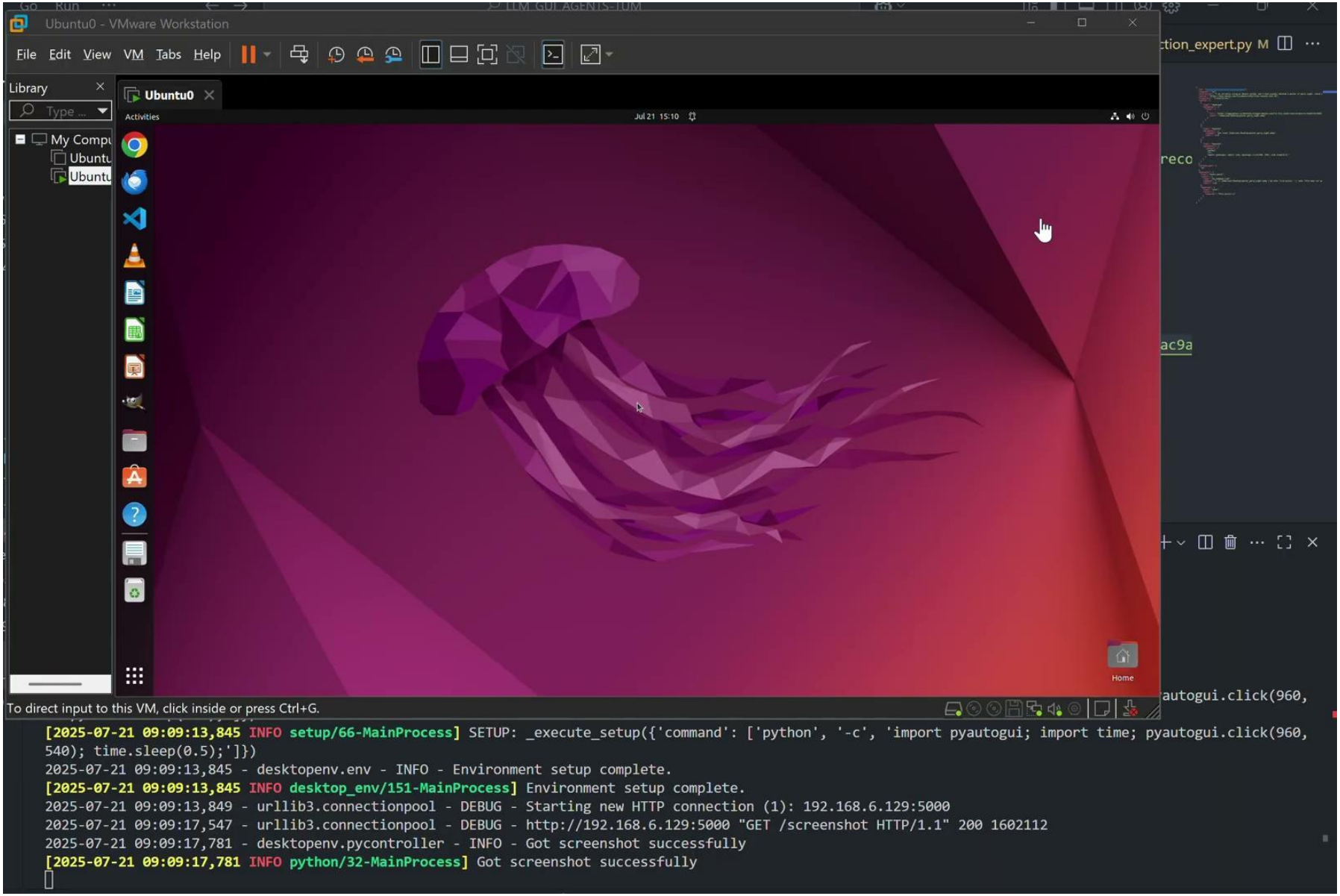
- Adapt to the benchmark
- Isolated responsibilities
- Links between experts

OSWorld Benchmark: Results

	OSWorld (5 tasks per section)				
	OS	**Chrome**	**Multitask**	**Libreoffice writer**	**Gimp**
Agent	3	4	0	2	1
				TOTAL	**10**/25

DEMO

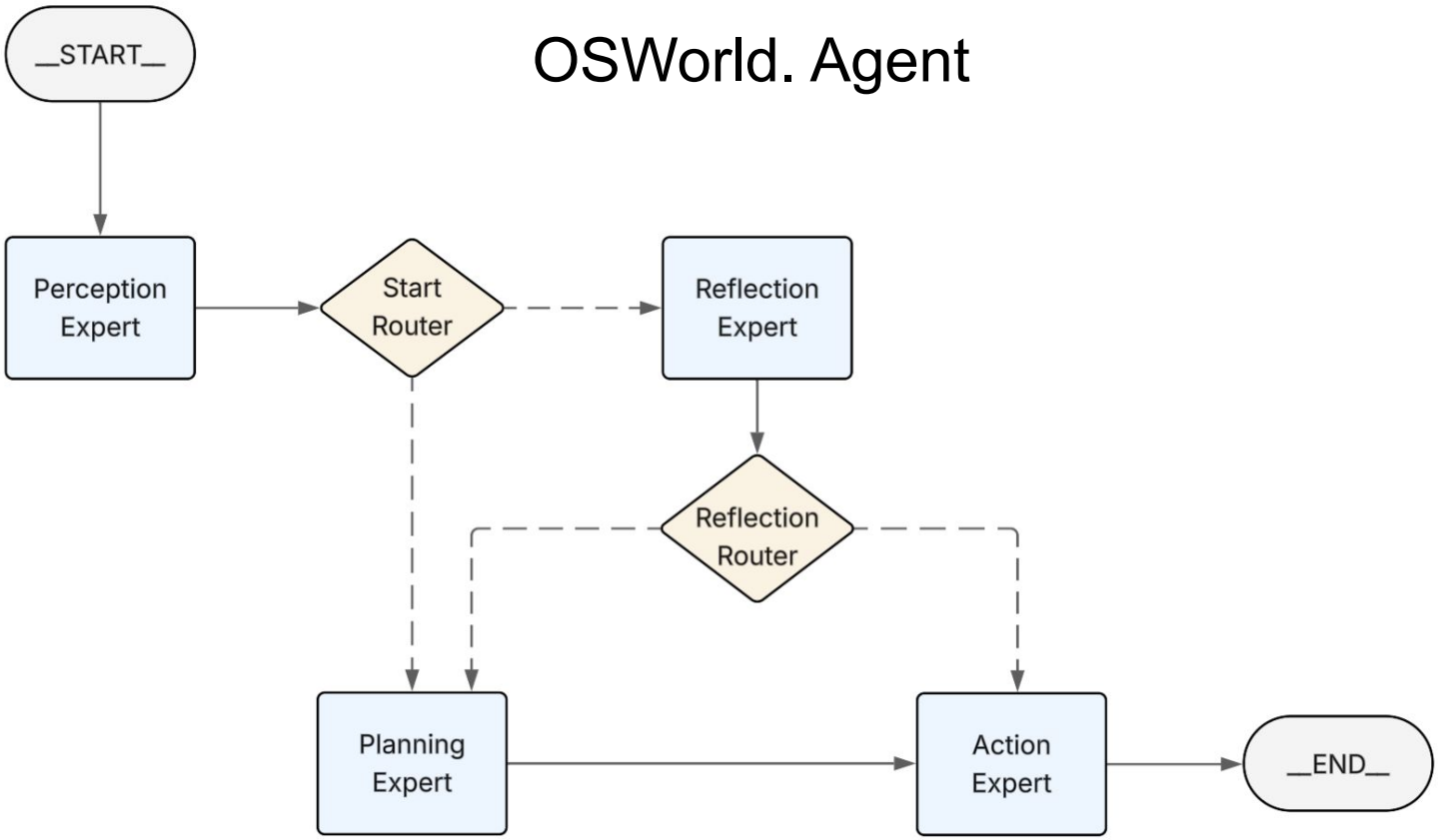
Demo


Ubuntu0 - VMware Workstation
File Edit View VM Tabs Help
Library
Ubuntu0
Activities
Jul 21 15:10
Home
To direct input to this VM, click inside or press Ctrl+G.
[2025-07-21 09:09:13,845 INFO setup/66-MainProcess] SETUP: _execute_setup({'command': ['python', '-c', 'import pyautogui; import time; pyautogui.click(960, 540); time.sleep(0.5);']})
2025-07-21 09:09:13,845 - desktopenv.env - INFO - Environment setup complete.
[2025-07-21 09:09:13,845 INFO desktop_env/151-MainProcess] Environment setup complete.
2025-07-21 09:09:13,849 - urllib3.connectionpool - DEBUG - Starting new HTTP connection (1): 192.168.6.129:5000
2025-07-21 09:09:17,547 - urllib3.connectionpool - DEBUG - http://192.168.6.129:5000 "GET /screenshot HTTP/1.1" 200 1602112
2025-07-21 09:09:17,781 - desktopenv.pycontroller - INFO - Got screenshot successfully
[2025-07-21 09:09:17,781 INFO python/32-MainProcess] Got screenshot successfully

Future Work

- Memory System
- Error Handling Expert
- Native GUI Agent

Q&A


OSWorld. Agent
__START__
Perception Expert
Start Router
Reflection Expert
Reflection Router
Planning Expert
Action Expert
__END__