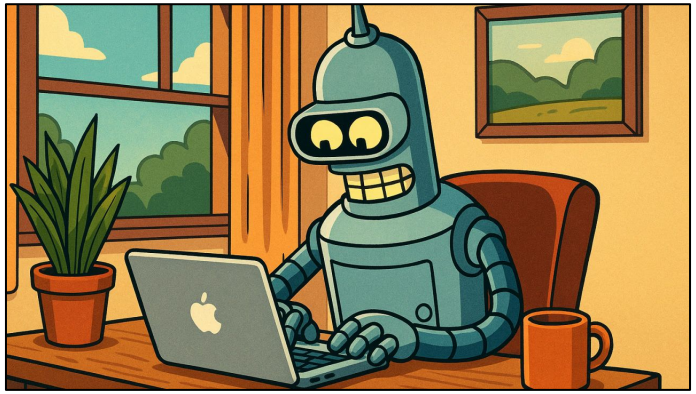


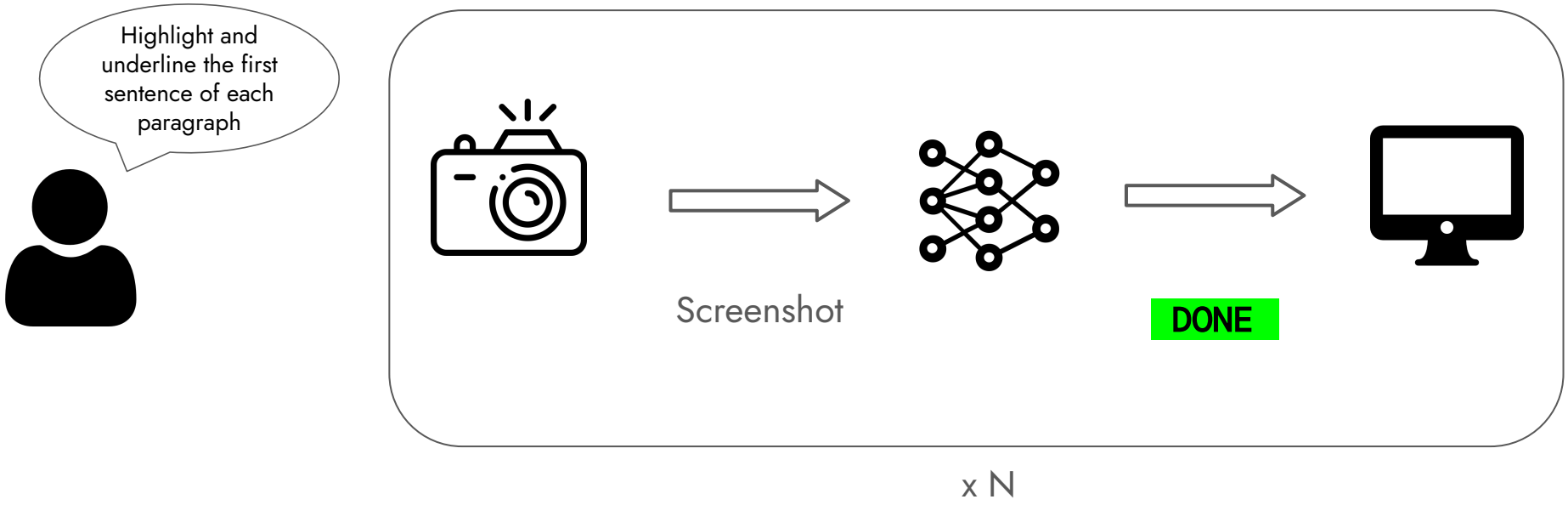
OSWorld-Human

Benchmarking the Efficiency of Computer-Use Agents

https://www.bytebot.ai/blog/the-age-of-the-desktop-agent-is-here

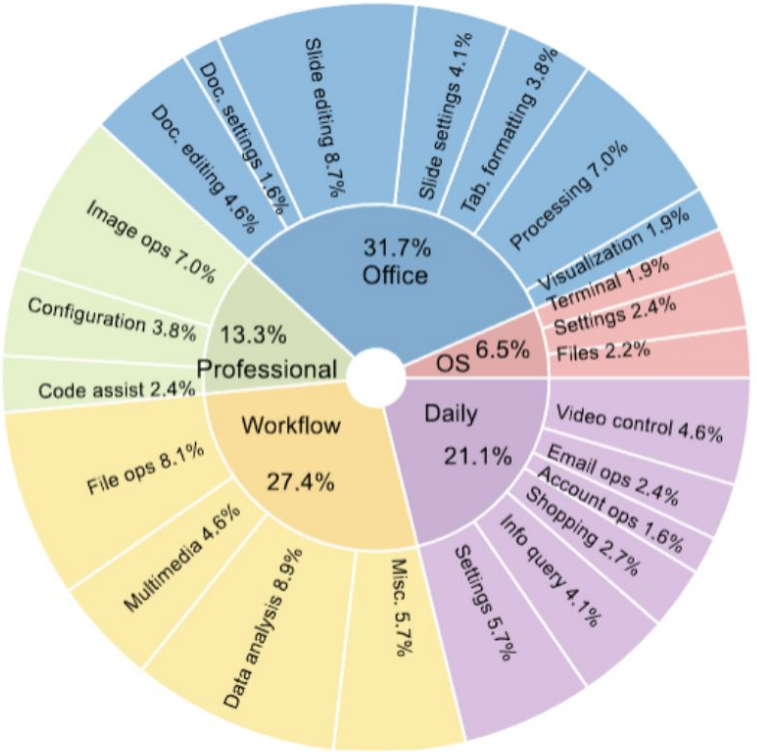
Reyna Abhyankar*, Qi Qi*, Yiying Zhang

Computer-Use Agent (CUA) Lifecycle



2024 OSWorld Benchmark

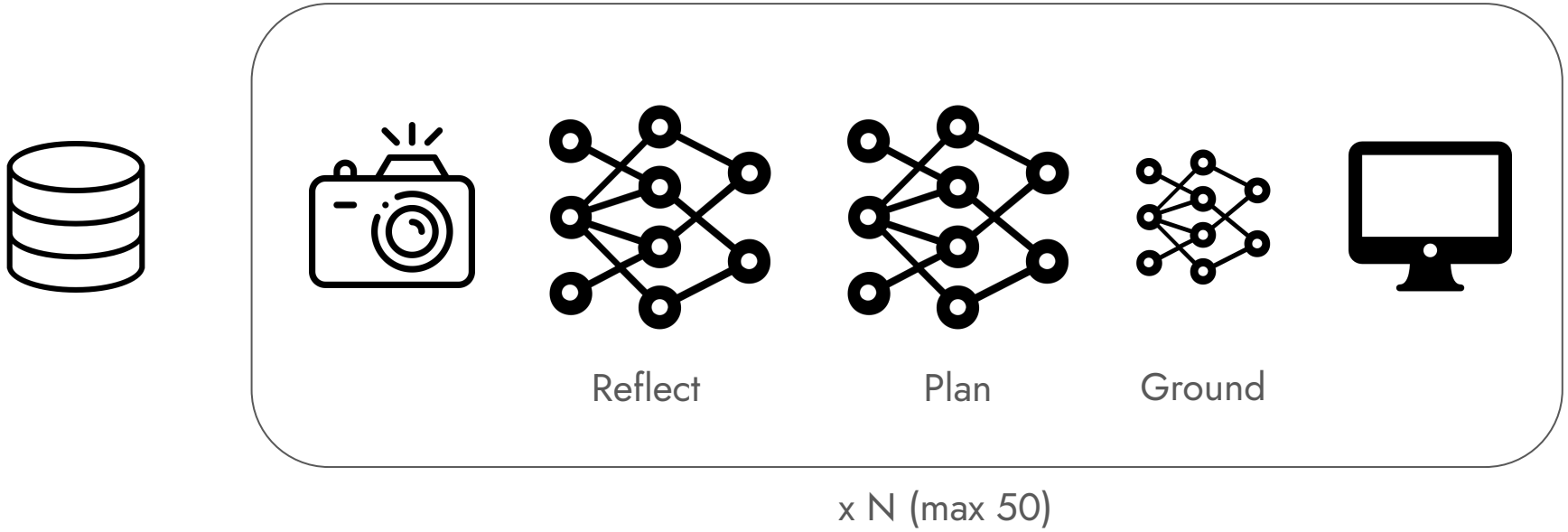
- 369 tasks
- Variety of domains
  - Chromium
  - LibreOffice Suite
  - VSCode
  - GIMP
  - VLC
  - Thunderbird
  - Terminal
- GPT-4: **12.2%**



https://os-world.github.io/

41.4%

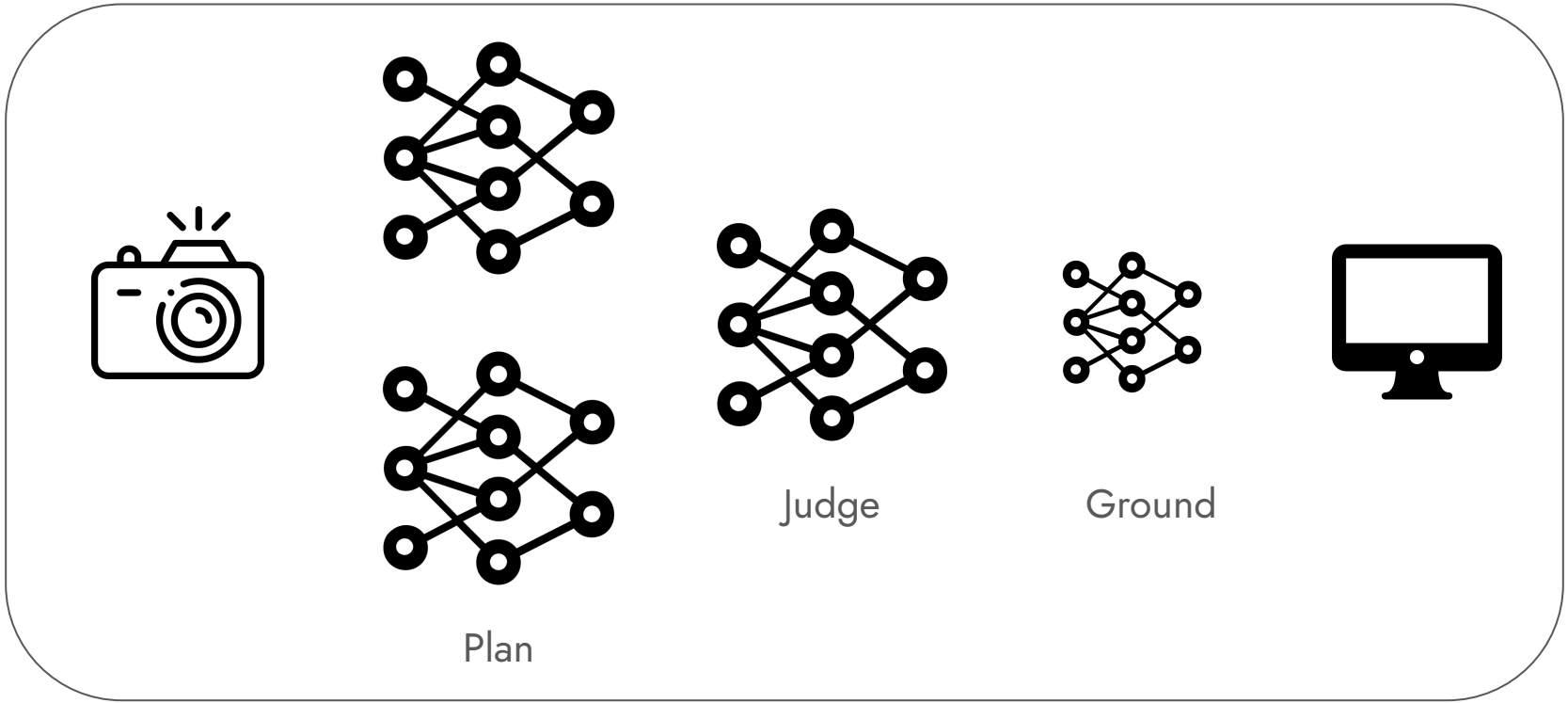
More Complex Agents: Agent S2



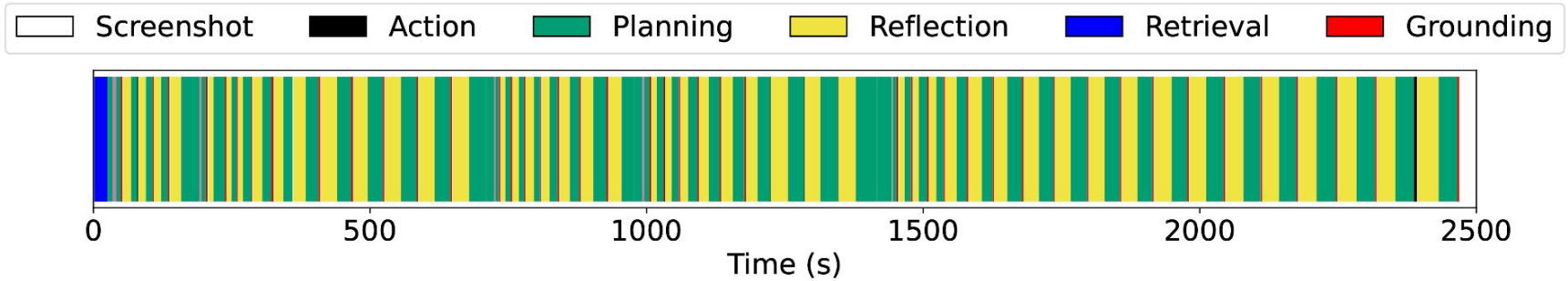
...

48.6%

More Complex Agents: GTA-1

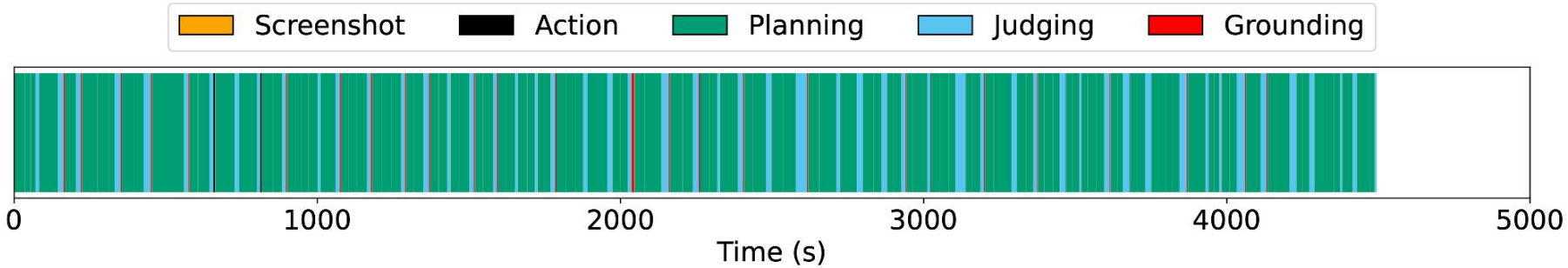


Tasks Take **Minutes!**



Agent S2

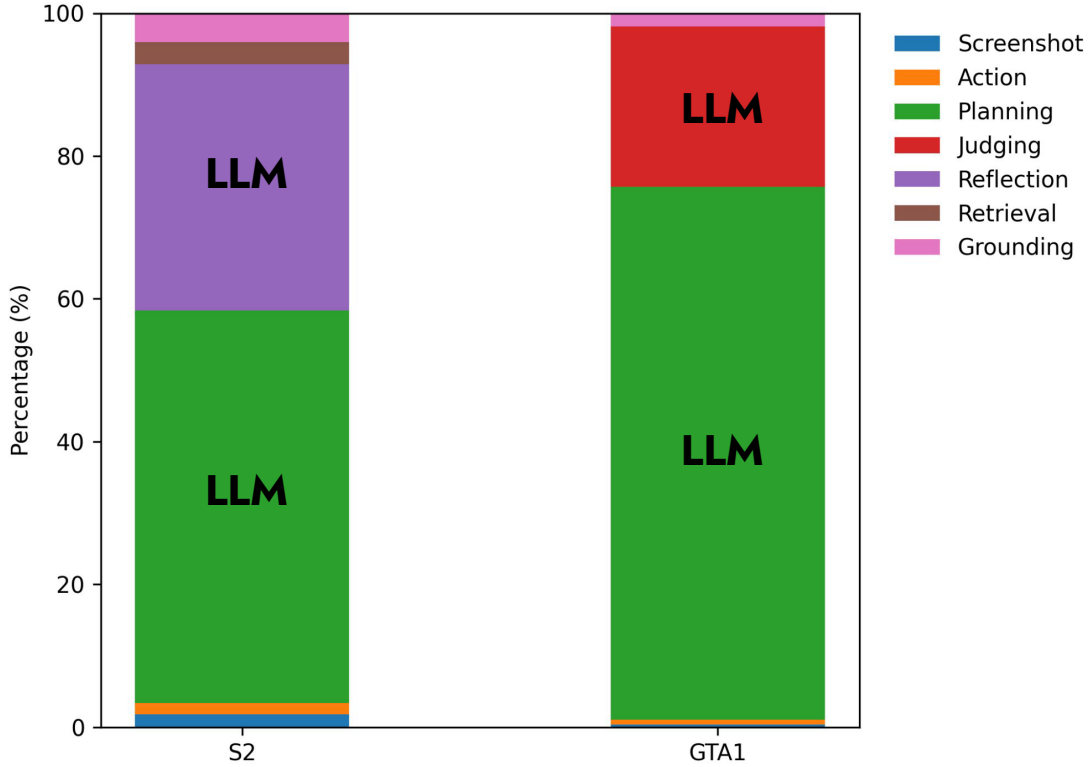
Tasks Take **Minutes!**



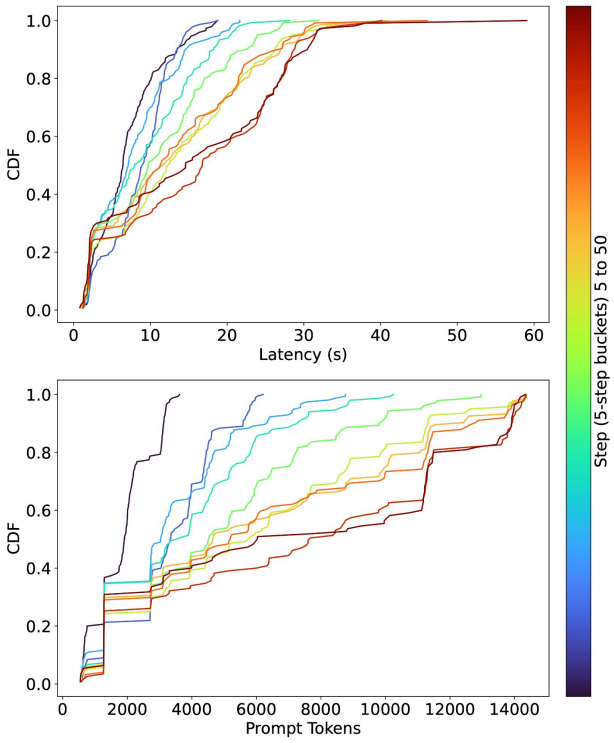
GTA-1

Breakdown by Stage

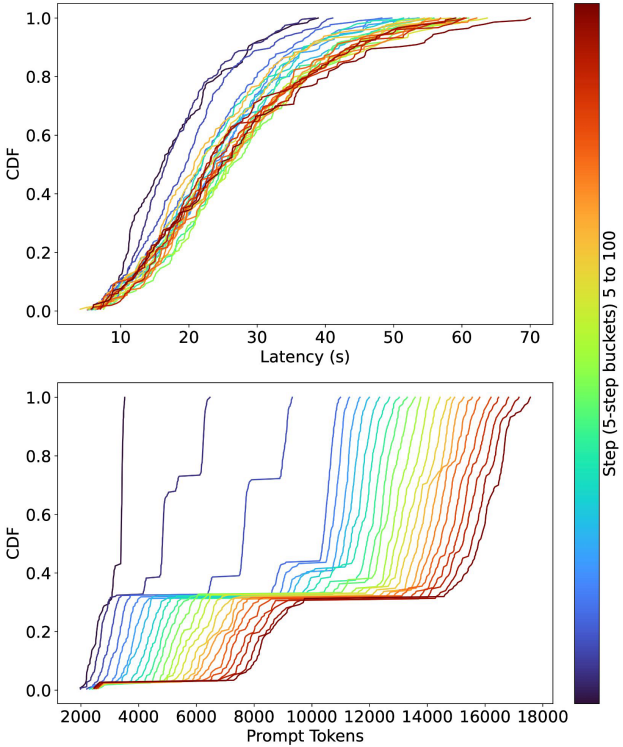
- p50 is 10-15k uncached prompt tokens!
- Translates to 20-30s **per-step**
- Planning consumes more prompt tokens than Reflection/Judging



Prompt Tokens Accumulate!

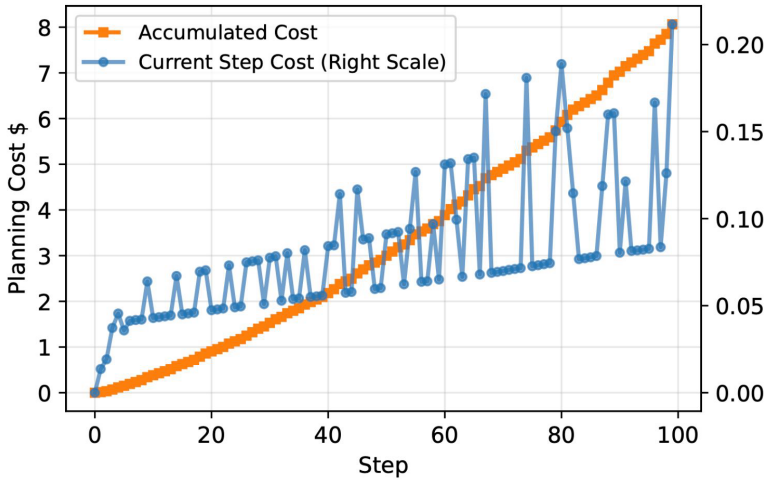


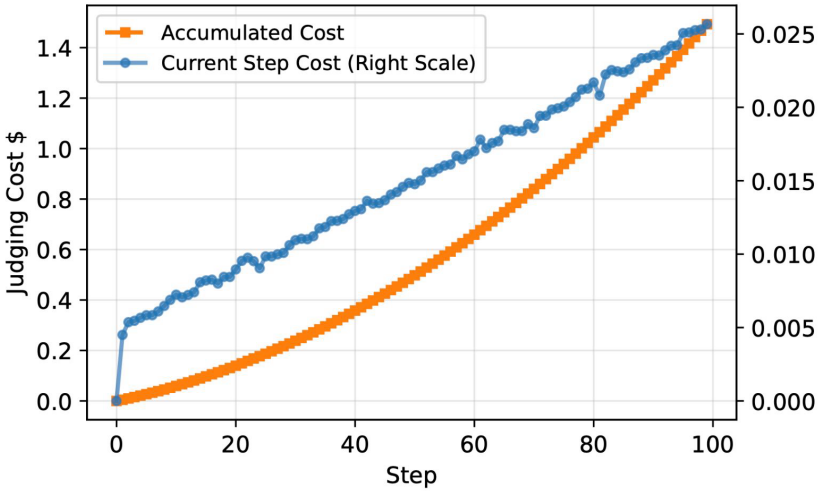
Agent S2



GTA-1

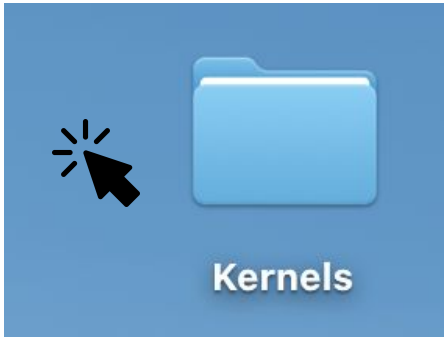
Cost ($) Accumulates Too!

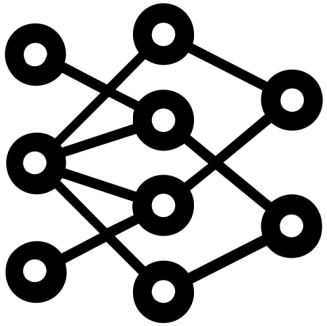




GTA-1

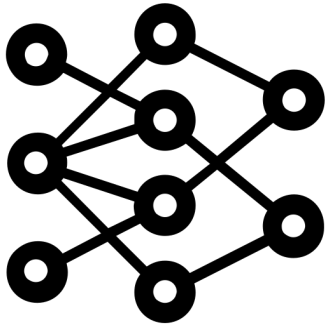
Failure Analysis: Looping





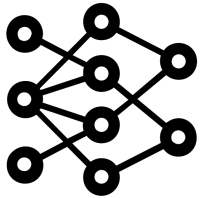
Reflection

"Wrong click location"



Plan

"Open the blue folder."



Ground

`CLICK(20, 25)`

For tasks that fail in 50+ steps,
66% of steps are wasted on looping

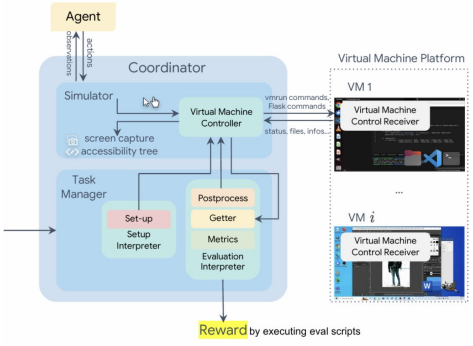
x N

None of this is captured by existing benchmarks

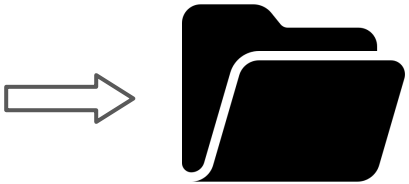
So we built one.

OSWorld-**Human**

- Recorded each step for all 369 tasks
- Re-scored agents according to a **Weighted Efficiency Score (WES)**
- Available open-source at https://github.com/WukLab/osworld-human



Agent vendor runs OS-World



Outputs result folder: **41.4%**

```
python score.py
       --result-path /x/y/z
       --max-steps-scoring 50
```

WES: **15.6%**

Scoring Efficiency

avg (reward)

Scoring Efficiency: Weighted Reward

$$\text{avg}\left(\text{reward} \cdot \frac{\text{human steps}}{\text{agent steps}}\right)$$

Scoring Efficiency: Global Penalty

$$\text{avg}\left(\text{reward} \cdot \frac{\text{human steps}}{\text{agent steps}}\right)\left(1 - \frac{\text{avg steps in } \textbf{failures}}{\text{max steps}}\right)$$

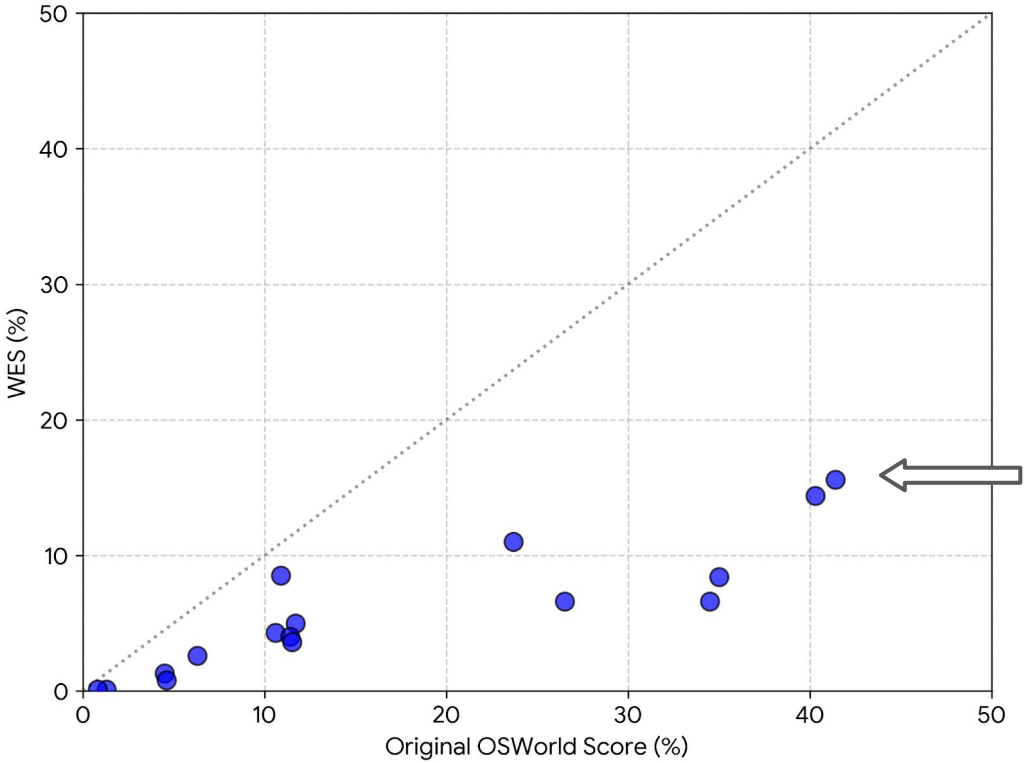
Weighted Efficiency Score (WES)

$$\frac{1}{n}\sum_{t}^{n} r_t \cdot \frac{t_{human}}{t_{agent}} \cdot \left(1 - \frac{\bar{t}_{fail}}{S}\right)$$

Weighted reward

Global penalty

Stacking up against OSWorld

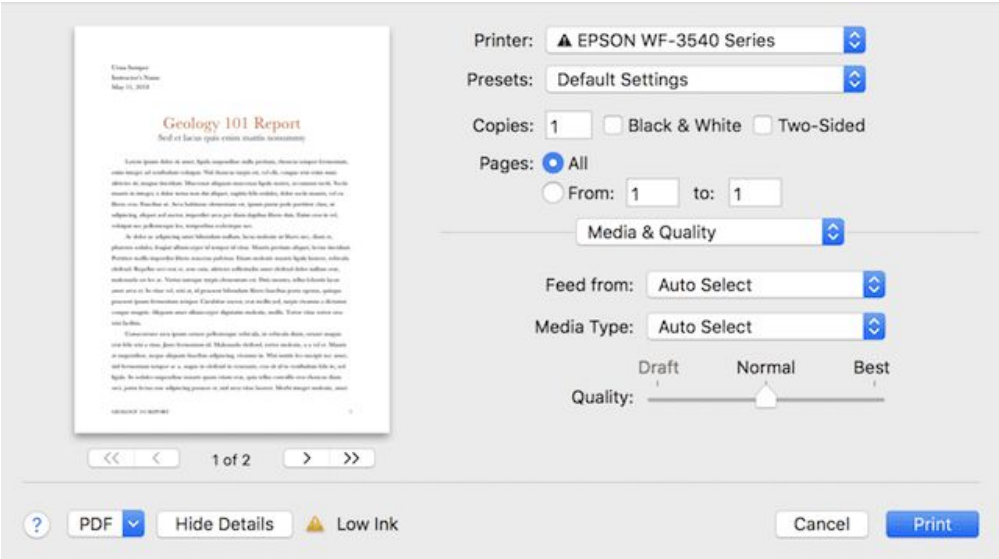


Agent S2

41.4% -> 15.6%

Grouped Actions

Task: Print this geology report in black & white.



Plan: Click “black & white” checkbox and then print.

⇩

`CLICK (19, 90)`

⇩

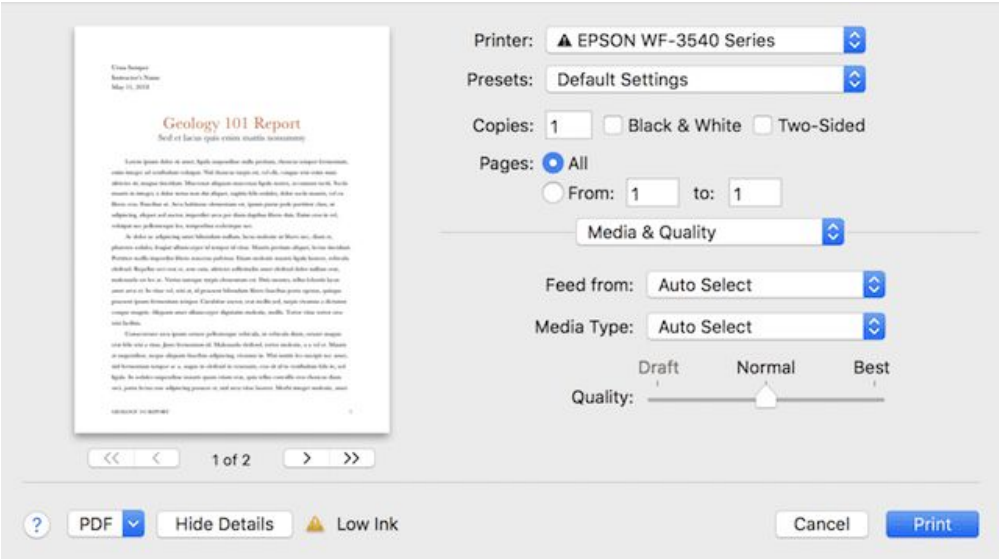
Plan: Click print.

⇩

`CLICK (20, 01)`

Grouped Actions

Task: Print this geology report in black & white.



Plan: Click “black & white” checkbox and then print.

⇩

```
CLICK (19, 90)

CLICK (20, 01)
```

Grouped Actions

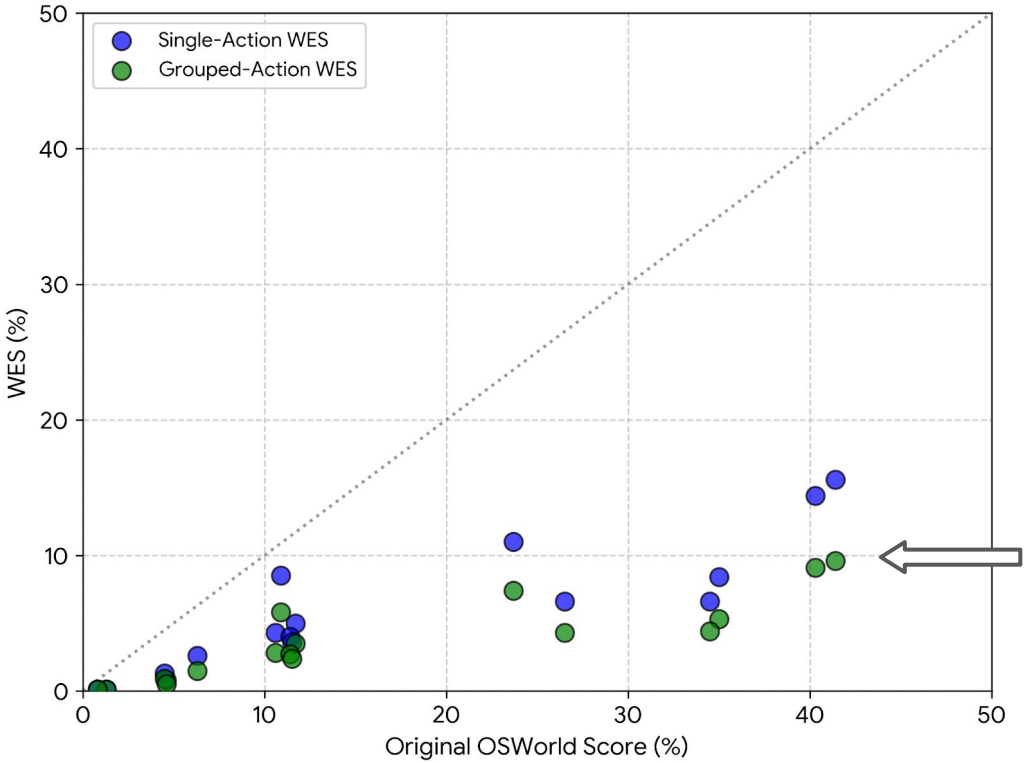
```
"single-action": [
  "`CLICK` on cell J2",
  "`TYPING` '=B2-C2-D2-SUM(F2:H2)'",
  "`CLICK` format as currency icon",
  "`MOVE_TO` bottom right corner of the cell J2`",
  "`DRAG_TO` bottom right corner of the cell J10",
  "`CLICK` on + to left of sheet1",
  "`TYPING` 'Year_Profit'",
  "`PRESS` enter",
  "`TYPING` '=$Sheet1.A2&\"_\"&$Sheet1.J2'",
  "`MOVE_TO` bottom right corner of the cell A2`",
  "`DRAG_TO` bottom right corner of the cell A10"
],
```

11 steps

```
"grouped-action": [
  [
    "`CLICK` on cell J2",
    "`TYPING` '=B2-C2-D2-SUM(F2:H2)'",
    "`CLICK` format as currency icon",
    "`MOVE_TO` bottom right corner of the cell J2`",
    "`DRAG_TO` bottom right corner of the cell J10",
    "`CLICK` on + to left of sheet1"
  ],
  [
    "`TYPING` 'Year_Profit'",
    "`PRESS` enter",
    "`TYPING` '=$Sheet1.A2&\"_\"&$Sheet1.J2'",
    "`MOVE_TO` bottom right corner of the cell A2`",
    "`DRAG_TO` bottom right corner of the cell A10"
  ]
]
```

2 steps

Stacking up against OSWorld



Agent S2

15.6% -> 9.6%

Conclusion

- OSWorld-Human: a benchmark that measures efficiency **and** effectiveness
  - Preserves relative ordering of OSWorld
  - Performance drops 41.4% → 9.6%
  - Best agents take 2.7-4.3x more steps than necessary
- What can we do about it?
  - Action Grouping
  - Efficient Rollback
  - Grounding Model Post-Training
  - History Compression
  - Improvements in LLM Serving
- Next steps: **benchmark your latest agents!**

https://github.com/WukLab/osworld-human