AdaptAgent: Adapting Multimodal Web Agents with Few-Shot Learning from Human Demonstrations

Gaurav Verma, Rachneet Kaur, Nishan Srishankar
Zhen Zeng, Tucker Balch, Manuela Veloso

Georgia Institute of Technology

JPMorgan Chase & Co.
AI Research

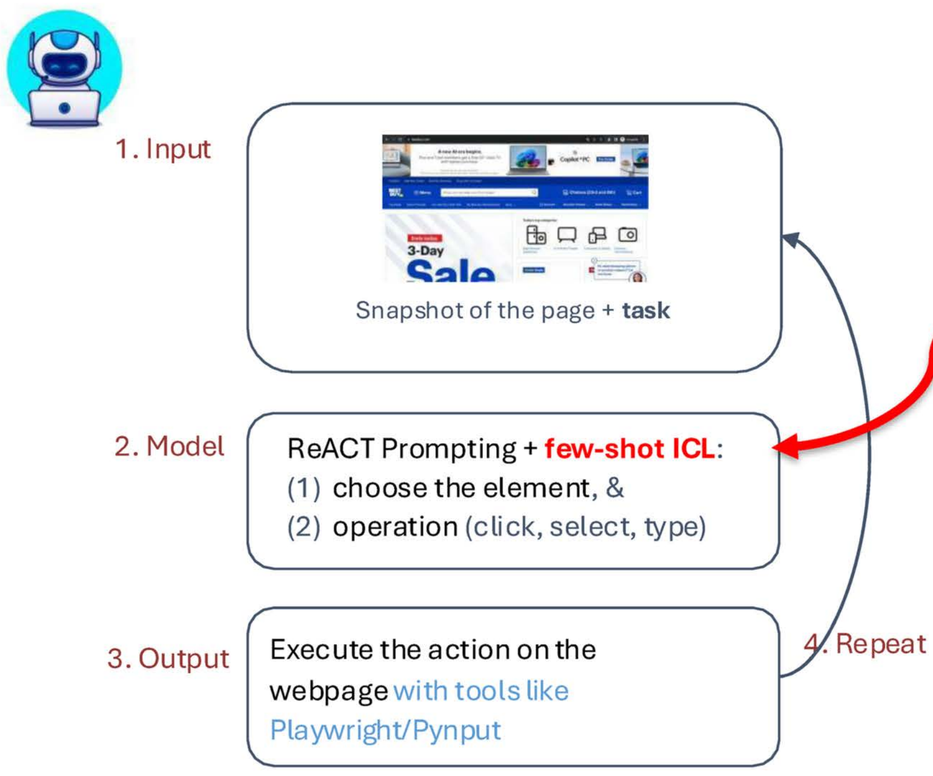
NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024 Workshop on Adaptive Foundation Models

How can we efficiently adapt multimodal web agents to work on unseen GUIs/webpages – including proprietary software and tasks?

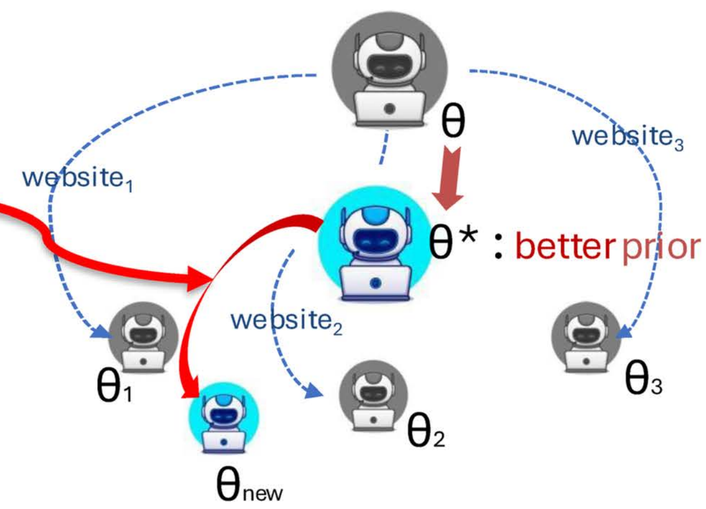
We propose the **AdaptAgent** framework!

***Upto 2 human demonstrations** to adapt MLLMs like GPT-4o (proprietary) using **in-context learning** and MLLMs like CogAgent (open-weights) using **meta-learning**.*

Including few-shot human demonstration with **ICL** for agents based on **proprietary** LLMs (SeeAct + 1-ICMD w/ GPT-4o)

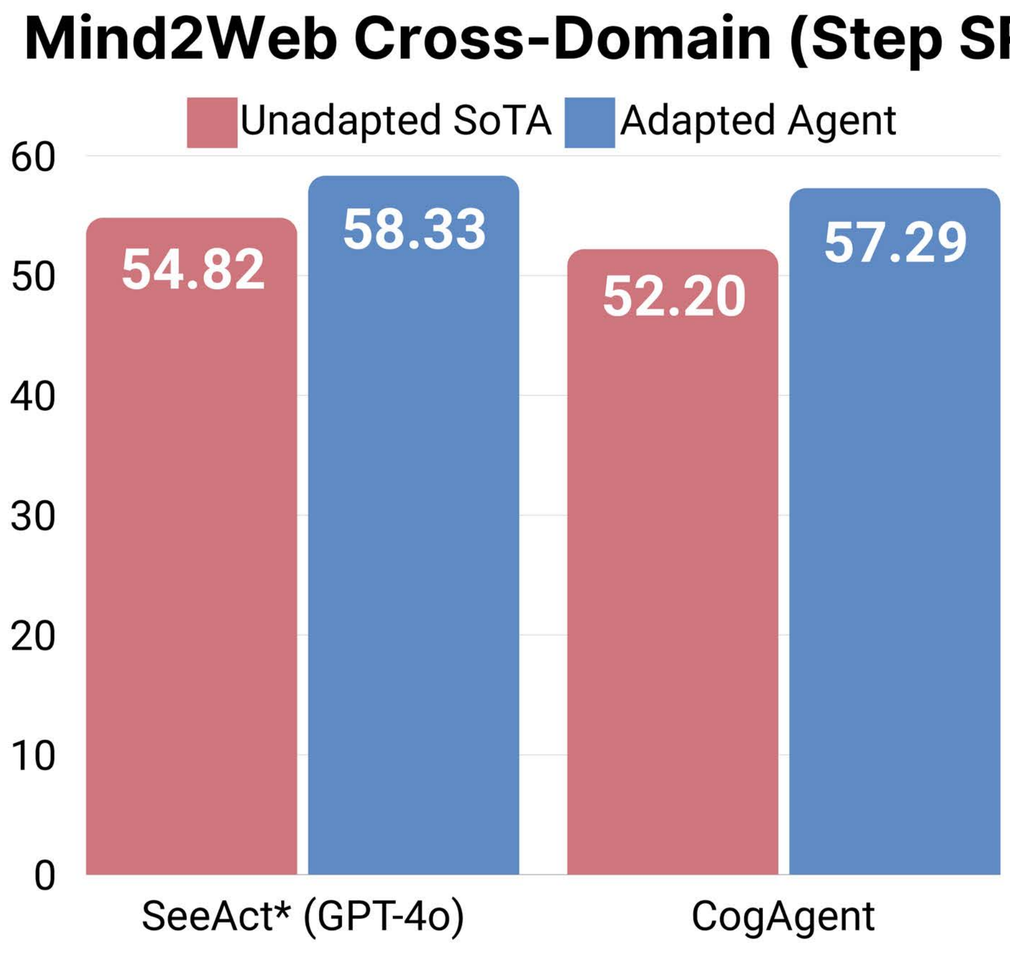


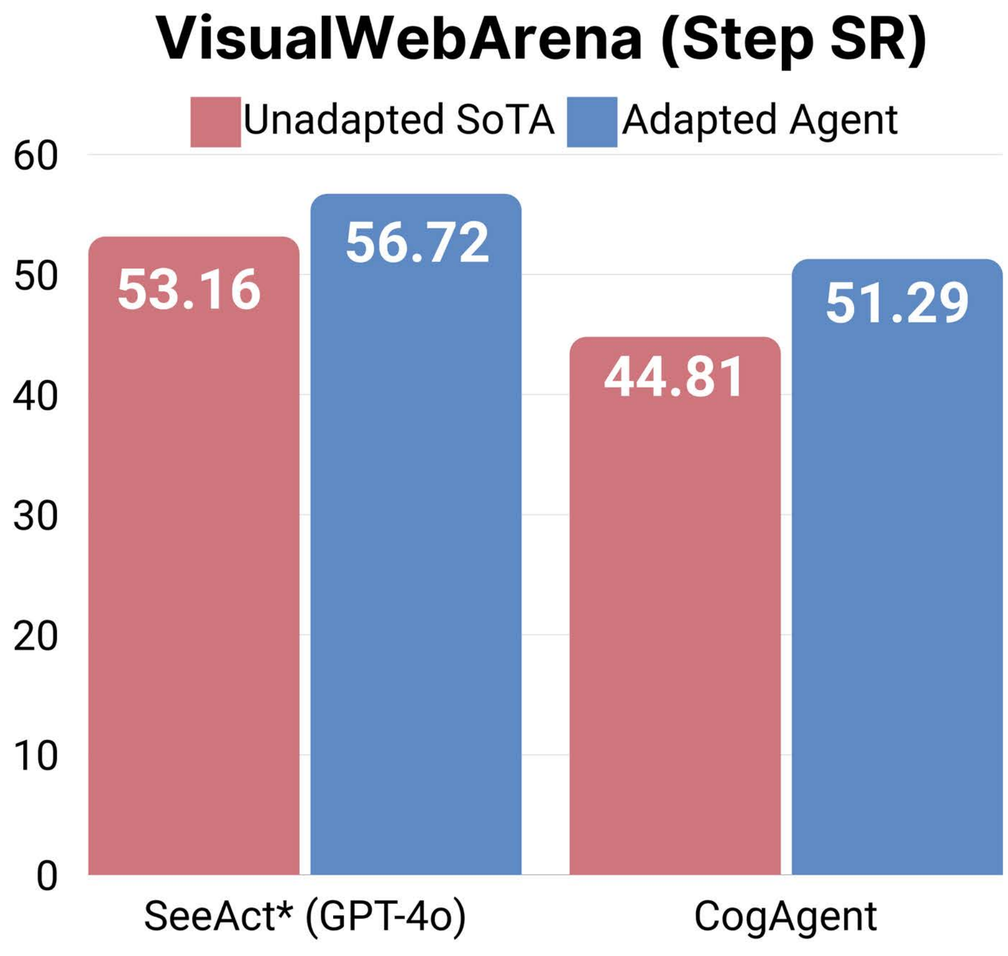
Meta-learning for few-shot adaptation of agents based on **open-weights** LLMs (CogAgent-FOMAML)



1. Meta-learning is used to learn a **better prior** ($\theta \rightarrow \theta^*$)
2. Human demonstrations ($n = 2$) are used to adapt θ^* to θ_{new}

Evaluation on benchmarks like Mind2Web and VisualWebArena: our framework boosts task success rate by 3.36% to 7.21% over non-adapted state-of-the-art models, corresponding to a relative increase of 21.03% to 65.75%.





- Multimodal human demonstrations are more effective than text-only demonstrations
- Increasing the number of in-context learning examples helps, but with diminishing gains
- Data selection strategy during meta-learning affects the generalization of the adapted agent