
AGENT WORKFLOW MEMORY

Zora Zhiruo Wang [♣] Jiayuan Mao [♠] Daniel Fried [♣] Graham Neubig [♣]

[♣] Carnegie Mellon University [♠] Massachusetts Institute of Technology

ABSTRACT

Despite the potential of language model-based agents to solve real-world tasks such as web navigation, current methods still struggle with long-horizon tasks with complex action trajectories. In contrast, humans can flexibly solve complex tasks by learning reusable task workflows from past experiences and using them to guide future actions. To build agents that can similarly benefit from this process, we introduce Agent Workflow Memory (AWM), a method for inducing commonly reused routines, i.e., *workflows*, and selectively providing workflows to the agent to guide subsequent generations. AWM flexibly applies to both *offline* and *online* scenarios, where agents induce workflows from training examples beforehand or from test queries on the fly. We experiment on two major web navigation benchmarks — Mind2Web and WebArena — that collectively cover 1000+ tasks from 200+ domains across travel, shopping, and social media, among others. AWM substantially improves the baseline results by 24.6% and 51.1% relative success rate on Mind2Web and WebArena while reducing the number of steps taken to solve WebArena tasks successfully. Furthermore, online AWM robustly generalizes in cross-task, website, and domain evaluations, surpassing baselines from 8.9 to 14.0 absolute points as train-test task distribution gaps widen.

1 INTRODUCTION

Language model (LM) based agents are rapidly improving, and are now able to tackle digital tasks such as navigating the web (Zhou et al., 2024; Deng et al., 2023) or operating mobile apps (Rawles et al., 2023; 2024). Current agents mostly integrate a fixed set of given examples via training (Fu et al., 2024; Murty et al., 2024) or in-context learning (Zheng et al., 2024). This allows them to perform well on action sequences similar to those presented in these examples, but results in a lack of robustness to changes in task contexts or environments (Deng et al., 2023). Essentially, they fail to grasp the key to disentangling increasingly complex tasks — to extract and learn reusable task workflows shared across similar tasks and environments (Yu et al., 2023; Wang et al., 2024). Moreover, as agents solve each task separately, they do not learn from past successes and failures, and are therefore unable to adapt over time (Yoran et al., 2024).

Motivated by how humans abstract common task routines from past experiences and apply such knowledge to guide future activities (Chi et al., 1981; 2014), we propose agent workflow memory (AWM) to realize a similar mechanism in agents. AWM induces workflows from agent trajectories by extracting reusable routines, and then integrates these workflows into agent memory to guide future task-solving processes. Each workflow represents a goal with a common routine extracted from available action trajectories, which allows it to effectively capture the most essential



Figure 1: AWM enables agents to continuously induce and apply workflows to improve performance, compared to stagnant baselines. We show results by AWM on the WebArena map split as an example.

and reusable skills agents need to acquire to successfully solve increasingly complex tasks. As an example, Figure 1 shows workflows induced by AWM on the WebArena map test split of the benchmark (Zhou et al., 2024). AWM starts with a basic set of built-in actions and solves new tasks in a

streaming manner, continuously inducing workflows from the task at hand, e.g., learning to “find a place by its name” from the first few examples. Moreover, AWM continues to build more complex workflows on top of new experiences and previously acquired workflows. For example, the “find a place by its name” workflow, once induced, effectively serves as a subgoal to build a more complex workflow “get the zip code of a place.” Such continual learning mechanisms create a snowball effect to induce and apply increasingly complex workflows while expanding the agent memory, often yielding a substantial performance gap over a vanilla agent that does not adapt. This gap over the baseline rises as high as 22.5 points on WebArena after rolling over only tens of examples (as shown by Figure 1).

AWM readily operates in both *offline* and *online* scenarios, where annotated examples are either available or non-existent. When high-quality annotated examples are available for a task, AWM operating in an *offline* fashion can extract reusable workflows from these canonical examples and integrate them into memory to assist test-time inference. Even if no annotated examples exist, AWM *online* can also run in an *supervision-free setting*, where it iteratively induces workflows from self-generated past predictions that are judged correct by an evaluator module.

We evaluate AWM on two agent web navigation benchmarks: WebArena, which provides rigorous execution-based evaluation (Zhou et al., 2024), and Mind2Web, which emphasizes broad tasks and domain coverage (Deng et al., 2023). On WebArena, AWM improves over the top published autonomous method (Drouin et al., 2024) by 51.1% relative success rate, and even outperforms methods augmented with human expert written workflows (Sodhi et al., 2023) by 7.9%. On Mind2Web, AWM effectively improves the cross-task results by 24.6% in relative step-wise success rate.

We further demonstrate the generalizability of AWM on both datasets. On WebArena, we create a cross-template subset where each example is instantiated from different task templates. AWM still consistently surpasses all baseline approaches, demonstrating its reliable cross-task workflow adaptability. On Mind2Web, we evaluate AWM on the cross-website and cross-domain test splits to examine its domain generality, where it scores 8.9–14.0 absolute points higher over baseline, and the margins become more substantial as the train-test distribution gap widens. Both results show the superior generalization of AWM across tasks, websites, and domains.

REFERENCES

- Micheline TH Chi, Paul J Feltovich, and Robert Glaser. Categorization and representation of physics problems by experts and novices. *Cognitive science*, 5(2):121–152, 1981.
- Micheline TH Chi, Robert Glaser, and Marshall J Farr. *The nature of expertise*. Psychology Press, 2014.
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=kiYqbO3wqw>.
- Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, et al. Workarena: How capable are web agents at solving common knowledge work tasks? *arXiv preprint arXiv:2403.07718*, 2024.
- Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. Autoguide: Automated generation and selection of state-aware guidelines for large language model agents. *arXiv preprint arXiv:2403.08978*, 2024.
- Shikhar Murty, Christopher Manning, Peter Shaw, Mandar Joshi, and Kenton Lee. Bagel: Bootstrapping agents by guiding exploration with language. *arXiv preprint arXiv:2403.08140*, 2024.
- Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy P Lillicrap. Androidinthewild: A large-scale dataset for android device control. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=j4b3l5kO1l>.

-
- Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*, 2024.
- Paloma Sodhi, SRK Branavan, and Ryan McDonald. Heap: Hierarchical policies for web actions using llms. *arXiv preprint arXiv:2310.03720*, 2023.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=ehfRiF0R3a>.
- Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. Assistantbench: Can web agents solve realistic and time-consuming tasks? *arXiv preprint arXiv:2407.15711*, 2024.
- Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. Language to rewards for robotic skill synthesis. *arXiv preprint arXiv:2306.08647*, 2023.
- Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. Synapse: Trajectory-as-exemplar prompting with memory for computer control. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=Pc8AU1aF5e>.
- Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=oKn9c6ytLx>.