Towards Unified Interfaces for Generalist Agent In Diverse Environments

Yueqi Song

CMU-CS-25-120

July 2025

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Graham Neubig, Advisor
Daniel Fried

Submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science..

Keywords: Agent, Reasoning, Large Language Model

For everyone who kindly offered me support and encouragement.

Abstract

Recently, large language models (LLMs) have enabled agents that can perceive, reason, and act in increasingly complex environments. Yet today's agents remain constrained by the interfaces they rely on, hampering generalization. This master thesis advances the goal of a *unified agent framework*.

Examining web agents, we found that web browsing agents, though intuitive to humans as they simulate human behaviours by browsing the web, are less effective and efficient. Thus, we proposed an API-based web agent that calls APIs through code generation, and demonstrated superior performance compared to browsing agents. Building on this, we further proposed a hybrid web agent that could interleave API calling and web browsing, broadening the agent's interface and allowing it to operate more effectively and efficiently in diverse environments.

Beyond web agents, we aim to extend the unified interfaces to generalist agents across diverse environments. To this end, we curated a large-scale unified training dataset that spans coding, web tasks, and general agentic tasks. The agent trained on this dataset achieved state-of-the-art (SOTA) performance on benchmarks testing a variety of tasks, marking a step towards unified interface for generalist agents.

Alongside a unified framework, strong reasoning abilities are crucial for agents to make correct decisions, plan, and execute tasks based on users' goals. We thus introduced VisualPuzzles, a benchmark that could evaluate models' multimodal reasoning abilities in a knowledge-light environment, which could provide guidance on the future development of models with strong multimodal reasoning capabilities.

Last but not the least, to serve people around the world, agents need to understand and generate multilingual contents. Thus, we proposed and trained Pangea, a multilingual model that achieved SOTA results on multilingual benchmarks.

Together, these contributions pave a path *towards unified interfaces for generalist agent in diverse environments*, providing the conceptual, empirical, and engineering foundations for the next generation of generalist AI agents.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Chapter 2

Conclusion

Bibliography