LUCI: Multi-Application Orchestration Agent

by

Guna Sekhar Sai Harsha Lagudu

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2024 by the
Graduate Supervisory Committee:

Aviral Shrivastava, Chair
Amarsagar Reddy Ramapuram Matavalam
Vidya Chhabria

ARIZONA STATE UNIVERSITY

May 2024

ABSTRACT

Research in building agents by employing Large Language Models (LLMs) for computer control is expanding, aiming to create agents that can efficiently automate complex or repetitive computational tasks. Prior works showcased the potential of Large Language Models (LLMs) with in-context learning (ICL). However, they suffered from limited context length and poor generalization of the underlying models, which led to poor performance in long-horizon tasks, handling multiple applications and working across multiple domains. While initial work focused on extending the coding capabilities of LLMs to work with APIs to accomplish tasks, a new body of work focused on Graphical User Interface (GUI) manipulation has shown strong success in web and mobile application automation. In this work, I introduce LUCI: Large Language Model-assisted User Control Interface, a hierarchical, modular, and efficient framework to extend the capabilities of LLMs to automate GUIs. LUCI utilizes the reasoning capabilities of LLMs to decompose tasks into sub-tasks and recursively solve them. A key innovation is the application-centric approach which creates sub-tasks by first selecting the applications needed to solve the prompt. The GUI application is decomposed into a novel compressed Information-Action-Field (IAF) representation based on the underlying syntax tree. Furthermore, LUCI follows a modular structure allowing it to be extended to new platforms without any additional training as the underlying reasoning works on my IAF representations. These innovations alongside the 'ensemble of LLMs' structure allow LUCI to outperform previous supervised learning (SL), reinforcement learning (RL), and LLM approaches on Miniwob++, overcoming challenges such as limited context length, exemplar memory requirements, and human intervention for task adaptability. LUCI shows a 20% improvement over the state-of-the-art (SOTA) in GUI automation on the Mind2Web benchmark. When tested in a realistic setting with over 22 commonly used applications, LUCI achieves an 80% success rate in undertaking tasks that use a subset of these applications. I also note an over 70% success rate on unseen applications, which is a less than 5% drop as compared to the fine-tuned applications.

DEDICATION

To all looking to explore the bounds of human knowledge and have fun doing it.

ACKNOWLEDGMENTS

I would like to express my most sincere gratitude to my thesis advisor, Dr. Aviral Shrivastava, for his active and patient support throughout my Thesis. His encouragement and technical input have been indispensable in this endeavor. Specifically, his help throughout the publication and writing process has helped provide a clear and concise narrative.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Automation and assisted computer interaction have been a significant area of investment for researchers and industry professionals. Digital assistants such as Siri and Google Assistant Tulshan and Dhage (2019) have enabled the creation of the Internet of Things (IoT) devices and the home automation space. They also play a critical role in the accessibility domain, enabling users with disabilities Isyanto *et al.* (2020) to interact with computers and mobile devices. Today these virtual assistants are regularly used to automate a variety of simple tasks Li *et al.* (2017) such as setting alarms and reminders, controlling music playback, etc. However, they have always struggled with more complex tasks and require specific language based on keywords

Instruction: Create a powerpoint on Recycling with Q & A slide at the end and Add Image Recycling.svg to last slide from Images folder in Downloads

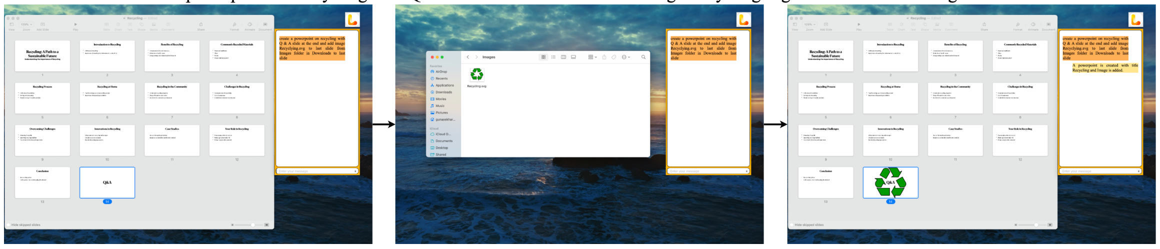

Figure 1.1: An illustrative execution trace of LUCI creating a presentation to satisfy the given instruction : ”Create a presentation on Recycling with Q & A slide at the end and Add Image Recycling.svg to last slide from Images folder in Downloads”. First LUCI opens Keynote Application and creates a presentation on Recycling. Then, open Images folder from Downloads directory and selects Recycling.svg file. Finally, LUCI adds the image to the last slide of the presentation. **This showcases LUCI’s ability to execute tasks involving multiple applications.**

to function correctly.

A key breakthrough in this domain was the advent of Large Language Models (LLMs), which worked directly with natural language and displayed strong reasoning and planning capabilities Ye *et al.* (2023). This allowed them to create a more capable and user-friendly interface for human-computer interaction (HCI) Desai *et al.* (2023). Applications such as ChatGPT and Google Gemini proved the promise of the underlying models which are now set to become the backbone of the aforementioned assistants. While LLMs proved to be effective in interacting with the user, their ability to interact with computer systems and use tools was still limited and needed to be augmented. The first approaches attempted to extend the coding capabilities of LLMs to work with APIs to accomplish tasks Yang *et al.* (2023); Schick *et al.* (2023). These systems adopted a Planner, Actor, and Reporter Dasgupta *et al.* (2023a) to ground the LLMs (restrict the response to relevant domains not part of the LLM's trained knowledge) and allow them to interact with the environment(computer systems). While they were initially successful, API-based LLM automation systems struggled with generalization across multiple applications due to the need for creating natural language instructions for the API of each application. These methods Ahn *et al.* (2022); Gao *et al.* (2023) tried to leverage *In-Context Learning* (ICL), to improve the generalization of the underlying models by forgoing training and instead placing relevant information in the context for prompts. However, the limited context lengths of these models led to poor performance in long-horizon tasks and handling multiple applications.

This led to the development of Graphical User Interface (GUI) LLM frameworksDeng *et al.* (2023a); Humphreys *et al.* (2022a). These systems focused on manipulating applications via the User Interface(UI) instead of relying on APIs which may or may not exist. The earliest versions used Reinforcement Learning (RL) to train an agent to mimic user clicks on inputs containing Hyper Text Markup Language(HTML) Document Object Model (DOM) elements. However, they struggled with with selection of relevant UI elements requiring supervised training. Gur et al Gur *et al.* (2023) demonstrated the difficulty in training LLMs with purely HTML due to the low information density and high noise in RAW HTML. Zeng et al Zheng *et al.* (2024a) incorporated the multi-modal learning capabilities of GPT4 to address the grounding of UI elements by proving the images of webpages to the model. This approach revolves around understanding the visual aspects of rendered web pages and generating precise plans in text format for various websites and tasks. However, this method did not scale beyond HTML and was hence limited to web applications. Additionally, these systems cannot utilize multiple applications or websites to accomplish a single task.

In this work, I introduce LLM-assisted User Control Interface (LUCI), a novel framework that extends the capabilities of LLMs to automate GUIs. LUCI is designed to enable LLMs to orchestrate multiple applications and execute complex tasks by interacting with the GUI. I accomplish this by adopting an application-centric approach to task planning where a *Tool Selector* element (Section. 3) selects the most relevant applications from a given set to solve a task. These are then used as the central focus when decomposing a given instruction into sub-tasks. Each sub-task is then mapped to a novel intermediate compressed structured representation based on Information-Action-Field (IAF) pairs via a*"UI Extractor"* (Section. 3) and a *'UI Selector"* (Section. 3). This structured representation allows the LLM to effectively interact with the GUI applications and execute the sub-tasks. The LUCI framework is designed to be modular, hierarchical, and OS-agnostic, enabling it to work seamlessly across both native and web interfaces. Additionally, I augment the performance of my conversational model with a *Task Verifier* (Section. 3) to filter redundant sub-tasks and improve efficiency. The limited scope of the Task Verifier allows it to focus on the relevance of a given sub-task, effectively reducing the generative task of the conversational model into a simpler decision task. The structure of these elements and how they interact with each other is shown in Figure 3.1.

When combined these components allow LUCI to solve complex multi-step and multi-application tasks across web and native interfaces without the need for additional multi-modal context. An example of this capability is shown in Figure 1.1, where LUCI creates a presentation on Recycling with a Q & A slide at the end and adds an image ”Recycling.svg” to the last slide from the Images folder in Downloads. This showcases LUCI's ability to execute tasks involving multiple applications.

The main contributions of LUCI can be summarized as:

1. An application-centric approach to task planning, where the selection of relevant applications is the basis of sub-task generation.

2. A modular OS-agnostic agent capable of functioning seamlessly across both native and web interfaces.

3. A novel tool selection mechanism is implemented to identify relevant tools for tasks involving multiple applications, enhancing adaptability and effectiveness.

4. A novel UI parser is designed to extract the web and desktop interfaces into a compressed and structured representation based on Information-Action-Field (IAF) pairs, thereby facilitating efficient orchestration by large language models (LLMs).

5. A hierarchical control structure within LUCI, enabling effective management of tasks across multiple applications, thereby empowering LLMs with comprehensive control in diverse environments.

The experimental results support my claim. I note that LUCI achieves a greater than 99% success rate on the MiniWoB++ benchmark. LUCI also achieves up to 31% improvement in Step SR and up to 24% improvement in OP. F1 Score over GPT4V on the Mind2Web benchmark. When I tested the generalization capability I noted a less than 5% drop in accuracy, indicating strong generalization. Finally, my experiments show that LUCI maintains a 75% average success rate when using 4 applications simultaneously. I believe that the application-centric design proposed in LUCI presents a promising direction for future research in GUI automation and multi-application orchestration.

Chapter 2

RELATED WORK

Building Agents with LLMs

Large language models (LLMs) have offered promising avenues for leveraging natural language in decision-making tasks. One approach involves enhancing LLMs with executable actions Mialon *et al.* (2023). Huang et al. Huang *et al.* (2022a) showed LLMs can plan and perform basic domestic activities by mapping embeddings to a predefined list of actions. However, their study lacks specificity for contemporary activities. Ahn et al. Ahn *et al.* (2022) introduced SayCan, grounding actions by multiplying candidate action probabilities with FLAN Wei *et al.* (2022) and the action's value function as an indicator of suitability. Huang et al. Huang *et al.* (2022b) extended SayCan with Inner Monologue, adding a feedback loop to select actions according to current state. However, Inner Monologue relies on a pre-trained, robot policy conditioned on language with limited flexibility, impeding generalization across various domains. Zeng et al. Zeng *et al.* (2022) combined LLMs conditioned on a robot policy along with a vision - language model (VLM) for open vocabulary pick-and-place tasks. Dasgupta et al. Dasgupta *et al.* (2023b) utilized Chinchilla Hoffmann *et al.* (2022) as a planner in PycoLab, requiring RL pre-training for their actor module to follow natural language instructions. Moreover, previous methods were restricted to the necessity for fine-tuning.

Recently, enhancing LLM effectiveness involves integrating them with APIs for utilizing external tools like information retrieval systems, code interpreters, and web browsers Glaese *et al.* (2022); Menick *et al.* (2022); Schick *et al.* (2023); Thoppilan *et al.* (2022). Integratings models with APIs involves five methods: 1) pre-training or fine tuning the model with API enabled examples Taylor *et al.* (2022); Schick *et al.* (2023). This approach has limited API space. 2) Another approach provides few examples on how to use APIs and use in-context learning of LLM Ahn *et al.* (2022); Gao *et al.* (2023); Lazaridou *et al.* (2022). This approach cannot accommodate numerous APIs due to limited context length. 3) Reinforcement learning with human feedback to enhance API usage Nakano *et al.* (2022). 4) Creating natural language documents or structured programs instructs the model Vemprala *et al.* (2023); Paranjape *et al.* (2023). 5) Lastly, using natural language documents with RLHF improves user feedback connectivity Liang *et al.* (2023). When using natural language documents the performance of LLM using a tool depends mainly on the documentation of an API Liang *et al.* (2023). So, API developers are required to uphold comprehensive and well-structured documentation. Also, API developers need to improve frequently based on cases where API fails to execute instructions provided by user and user feedback. However, these API tools require manual engineering and may have limited functionality. So, recent approaches shifted to agents working with Graphical User Interface.

Automated GUI Tasks

Pursuing the goal of interaction between humans and the computer, significant efforts have been dedicated to developing autonomous computer agents capable of understanding language instructions and efficiently carrying out tasks on a computer Pasupat *et al.* (2018); Gur *et al.* (2018); Furuta *et al.* (2024a); Liang *et al.* (2023). To evaluate the models for human-like computer interactions, MiniWoB++ extending the MiniWoB benchmark Shi *et al.* (2017); Liu *et al.* (2018), serves as a standard benchmark. Early researchers utilized reinforcement learning and imitation learning to solve MiniWoB++ challenges Liu *et al.* (2018); Gur *et al.* (2018); Jia *et al.* (2019); Gur *et al.* (2022), but reaching human-level performance necessitates a large amount of expert demonstration data (6,300 hours) Humphreys *et al.* (2022b).

Recent research Gur *et al.* (2023); Furuta *et al.* (2024a) proposes employing large language models (LLMs) to read HTML code and vision transformers Dosovitskiy *et al.* (2021); Hong *et al.* (2023) for extracting screenshot features, using a few-shot in-context method yielding promising results without prolonged RL training. Nonetheless, large volumes of expert demonstration data are still needed to fine-tune LLMs. WebGPT Nakano *et al.* (2022) and WebShop Yao *et al.* (2023) demonstrate LLMs automating web tasks with custom commands, but they are restricted in scope and don't address general computer tasks requiring keyboard and mouse inputs.

RCI Kim *et al.* (2023) achieves a 90.6% success rate in 54 MiniWoB++ tasks using recursive self-correction, yet its reliance on task-specific examples restricts generalization to new scenarios. In contrast, my proposed method works well without relying on self-correction. AdaPlanner Sun *et al.* (2023) achieved a 92.9% success rate in 53 tasks by leveraging environment feedback for self-correction, but faced similar generalization challenges as RCI. Pix2Act Shaw *et al.* (2023) addressed 59 MiniWoB++ tasks through tree search and BC, based on 1.3 million demonstrations. WebGUM Furuta *et al.* (2024b) fine-tunes Language Multimodal Model (LMM) with a huge multimodal corpus for web agents, allowing web agents to observe both HTML and the captured screenshot but require lot of expert demonstrations for fine-tuning. Synapse Zheng *et al.* (2024c) uses structured prompts with a LLM and achieves human level performance. However, the performance is largely dependent on the quality of examples passed through prompts. On the other hand, my proposed method uses few-shot in-context learning with structured prompts and structured representation of UI elements on a browser making the agent independent of the quality of examples.

Moreover, my approach excels in addressing open-domain tasks on a large scale.

Numerous ongoing initiatives are dedicated to the development of computer agents and benchmarks. Among the sophisticated benchmarks available, such as Mind2Web Deng *et al.* (2023b), Webshop Yao *et al.* (2023), and WebArena Zhou *et al.* (2023). WebArena Zhou *et al.* (2023) generates website simulations in a sandbox environment across four popular categories, replicating functionality and data from real-world equivalents. In contrast, Mind2Web Deng *et al.* (2023b) offers environments spanning diverse domains, websites, and various tasks extracted from live websites, complete with action trajectory annotations. So, I have chosen Mind2Web for the real-world evaluation. To solve Mind2Web benchmark, MindAct utilized an element-ranking model to extract the top-50 relevant elements as clean observations. It employs MCQ to recursively query the LLM for action selection from five alternative candidates until either all options are wrong or just one action is selected. While MCQ-formatted exemplars perform better than direct generation, MindAct frequently has trouble selecting the right element Deng *et al.* (2023b). Another approach involves use of language and vision transformer. Recent works in this works includes WebGUM Furuta *et al.* (2024b) and Gpt-4v Zheng *et al.* (2024b). WebGUM Furuta *et al.* (2024b) fine-tunes Language Multimodal Model (LMM) with a huge multimodal corpus for web agents, allowing web agents to observe both HTML and the captured screenshot but require lot of expert demonstrations for fine-tuning. Gpt-4v Zheng *et al.* (2024b) uses LMM's to observe both HTML and captured screenshot to generate actions but still lack sufficient visual perception abilities to serve as an effective agent.

Chapter 3

PROPOSED METHOD

In this work, my goal is to enable large language model to use various GUI tools to extend its capabilities. The primary architecture of LUCI comprises of 7 key components, namely

1. **GUI Tool Set**, which contains list of GUI tools authorized to be controlled by the LLM.

2. **Tool Selector**, which selects GUI tools from GUI tool set required to accomplish given user instruction.

3. **Conversational Model**, which is responsible for interacting with users and generating a solution outline based on GUI tool selected by the tool selector, user instruction along with its conversational context.

4. **Task Verifier**, which filters the redundant sub-tasks in the solution outline using action feedback.

5. **UI Extractor**, which extracts the information of UI elements from the selected GUI tool.

6. **UI Selector**, which selects appropriate UI elements from the list of UI elements generated by UI extractor for the given sub-task.

7. **Action Executor**, which performs action on selected UI elements based on type of UI element and return the action feedback.

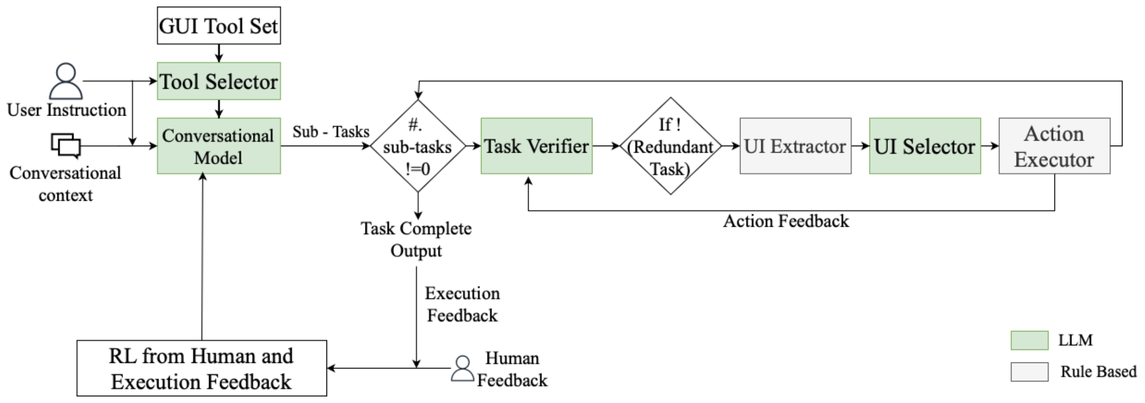

Figure 3.1: **Architecture of LUCI**. Given user instruction and the conversational context, the Tool Selector first selects a tool from the given GUI tool set and opens a GUI applications. Then, the conversational model generates a solution outline, which is a textual description of list of sub-tasks needed to solve the task. Next, Task Verifier filters redundant tasks in solution outline based on action feedback from previously executed tasks and future sub-tasks. Then, a rule-based UI Extractor extracts UI elements from GUI application. UI Selector selects appropriate UI elements from the list of UI elements generated by UI extractor for the given sub-task. Lastly, Action Executor performs action on selected UI element based on type of UI element and generates an action feedback.

The overall architecture of LUCI is shown in Figure 3.1. The primary process within this architecture is the LLM's capability to execute action in response to user instructions. This approach takes 4 inputs: large language model's parameters, represented as θ; GUI tool set, denoted as $\mathcal{G}$; the user instructions, designated as $\mathcal{I}$, along with the conversational context, referred to as $\mathcal{C}$. Utilizing these inputs, the LLM generates a set of actions, designated as $\mathcal{A}$, to execute and fulfill the user's instruction. This procedure can be defined as follows:

$$\mathcal{A} = LLM(\theta, \mathcal{G}, \mathcal{I}, \mathcal{C}) \tag{3.1}$$

Further, LUCI is employed with RLHF learning mechanism Ouyang *et al.* (2022) to improve the task planning and task verification. Here, in addition to human feedback, I also employ execution feedback, this combined feedback is denoted by $\mathcal{F}$. The Loss function of the learning mechanism is parameterised by,

$$\mathcal{L} = \mathcal{L}(LLM(\theta, \mathcal{G}, \mathcal{I}, \mathcal{C}), \mathcal{F}) \tag{3.2}$$

GUI Tool Set

The GUI tool set is a primary element in this framework, the tool set G = $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3.....\}$, which contains a collection of different GUI application names. The GUI tool set contains a list of all desktop application names (For instance, a weather application on your personal computer) that the agent is allowed to work on. In this work, I have used 22 GUI based desktop applications as shown in Table 1. In this work web applications are accessed through a browser like Safari, Google Chrome or Microsoft Edge, which gives more flexibility for navigation and using multiple GUI applications.

Tool Selector

The main objective of the tool selector is to select a most appropriate GUI tool from the GUI tool set that aligns with the given task requirements. The tool selector is a language model that uses in-context learning to select a GUI application $\mathcal{G}_s$ from the set of GUI tools $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3.....\}$ based on given user instruction I, as shown below. The output of the language model is grounded to application names in the GUI tool set.

The tool selector in this study facilitates the use of multiple GUI tools based on user instructions as shown in Example Query 2 in Figure 3.2. If a specified GUI tool is incapable of performing a task, ”Tool Not Found” is chosen as shown in Example

GUI Tool Set (G) : Calculator, Calendar, Keynote, Maps, Messages, Music, News, Numbers, Pages, Safari, Stocks, TextEdit, Weather, Script Editor, Terminal, Outlook.
Example Query 1: From the given list of GUI tools G, select a tool to solve the task " Send an email to johndoe@gmail.com".
Response: Outlook
Example Query 2: From the given list of GUI tools G, select a tool to solve the task "Weng earns $ 12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn? And write the entire proof."
Response: Calculator, Text Edit
Example Query 3: From the given list of GUI tools G, select a tool to solve the task " Draw a 3D model of cube with dimensions 3m x 3m x 3m".
Response: Tool Not Found
Example Query 4: From the given list of GUI tools G, select a tool to solve the task " Write a summary of the given paragraph, ... ".
Response: No tool required

Figure 3.2: Illustrative of examples of Tool Selector in selecting a GUI application from a given GUI Tool set. Example Query 1 shows the ability of Tool selector to select a GUI application. Example Query 2 shows the ability of Tool selector to select multiple GUI application. Example Query 3 shows the ability of Tool selector to notify the conversation model that given tasks cannot be performed with given GUI applications. Example Query 4 shows the ability of Tool selector to to notify the conversation model that given tasks can be solved without GUI application.

Query 3. For tasks that solely require language model capabilities, "No tool required" is selected as shown in Example Query 4. In cases of duplicate functionality, the user's specified tool takes precedence, or the tool is selected alphabetically or based on previous user selections.

Conversational Model

The conversational model is a LLM, which acts as a "brain" for this framework by understanding user intentions from the current instructions and past conversations to generate a set of sub-tasks (solution outline) needed to use the selected GUI application. To generate a Solution outline, I employed the Chain-of-thought (CoT), as it has shown significant improvement in performance across various tasks including arithmetic reasoning, commonsense reasoning, and symbolic reasoning Wei *et al.*

Example Query 5: What are the tasks to perform on Calculator and Text Edit for the given task “Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn? And write the entire proof”. Generate the tasks in the format ”Tool Required: True / False, Tool: tool Name, Task description: description of the task”.
Response:

1) Tool Required: False, Tool: Calculator, Task description: Open Calculator on your Macbook.
2) Tool Required: True, Tool: Calculator, Task description: Set the calculator to the ”Standard” mode”.
3) Tool Required: True, Tool: Calculator, Task description: Convert 50 minutes to hour by dividing 50 by 60 in the Calculator.
4) Tool Required: True, Tool: Calculator, Task description: Press the equals (=) key to calculate the result.
5) Tool Required: True, Tool: Calculator, Task description: Multiply number of minutes with money earned per hour ($12) into the Calculator.
6) Tool Required: True, Tool: Calculator, Task description: Press the equals (=) key to calculate the result.
7) Tool Required: True, Tool: Calculator, Task description: Observe the calculated result displayed on the Calculator screen.
8) Tool Required: False, Tool: Text Edit, Task description: Open Text Edit.
9) Tool Required: False, Tool: Text Edit, Task description: Generate a summary of the working.
10) Tool Required: True, Tool: Text Edit, Task description: Write the summary to Text Edit.

Figure 3.3: Illustrative example of solution outline from Conversation model to solve a task which involves integration of two applications, Calculator and Text Edit

(2023); Kojima *et al.* (2023). An example of a solution outline generated for using a calculator is shown below.

To facilitate the utilization of multiple GUI tools, I have specified a field called “Tool Name” for each sub-task along with sub-task description within the generated solution outline. The generated solution outline, consists of 3 types of sub-tasks:

1. Sub-tasks that contain description of opening an application. In this work, the GUI application is opened right after the tool selector selects an application name from the tool set and is not considered as a sub-task.

2. Sub-tasks that can be directly performed by the conversational model, such as text summarization (Sub-task 9 in the Example Query 5), will not trigger any action on GUI application.

3. Sub-tasks that contain action descriptions to be performed on GUI applicaations as seen in Sub-tasks 2-7 and 10 in the Example Query 5 in Figure 3.3.

From above categorization, to distinguish sub-tasks requiring GUI interaction, a boolean field ”Tool Required” is introduced in the Solution outline, preventing unnecessary execution of subsequent steps for tasks not involving tool interaction. The final output format of each sub-task in the solution outline is shown below:

Tool Required: True/False, **Tool**: GUI Application Name, **Task description**: Sub task description

When "No tool required" is chosen, the conversational model directly responds to user instructions, such as summarizing a text, and provides feedback in cases where "Not Found" is selected by the tool selector.

Task Verifier

Before executing each sub-task in the generated solution outline, it should be evaluated for 2 things to remove redundant tasks and improve efficiency:

1. Whether the sub-task is necessary or not based on future tasks.
2. Sub-task is already executed or not based on previous tasks.

In the response of Example Query 5, after converting 50 minutes to hours, there is no need to press "=" key, one can directly multiply this with the hourly earnings. Also, there might be a chance of pressing "=" key after dividing 50 by 60 in Sub-task 3 is valid, but doing so in Sub-task 4 leads to an error due to redundant division by 60. To prevent errors and redundant sub-tasks, I introduced Task verifier. The main aim of the Task verifier is to identify and remove the unnecessary sub-tasks in the solution outline.

Task verifier is a language model which takes current sub-task, future sub-tasks, and action feedback from previous sub-tasks to decide whether to proceed with execution or not by reasoning as shown in Example Query of Figure 3.4. The action feedback is generated through the analysis of interactions with user interface (UI) elements and the resulting changes in the user interface. This feedback mechanism informs the task verifier about actions performed in the past and reduces the execution of redundant actions in subsequent sub-tasks. At the end of execution, action feedback is combined to generate the execution feedback used in RLHF for model improvement.

Example Query 6 : Check whether the current task s3 is necessary for the objective I based on previously executed tasks {(s1; AF1), (s2; AF2)} and future tasks s4, s5, s6, s7. The answer format should be " Reason: Reason to execute sub-task , Answer: Yes/No".
Response: Reason: This task is required to convert 50 minutes to hours, Answer: Yes

Figure 3.4: Illustrative example of Task verifier in deciding whether a task is redundant or not by reasoning.

UI Extractor

I developed a rule-based UI element extractor to extract UI elements from desktop applications and its parameters like type of UI element, allowed actions, location, size, description, value (if any) from GUI tool. In any desktop application, the graphical interface is based on hierarchy / tree structure. I can use accessibility tools provided by Windows and MAC OS to extract this tree structure. In this tree based structure, I divide the UI elements into 3 categories :

- **Information Elements (I)**: UI elements that contain only Information. Example: $< p >, < h1 > - < h6 >, < span >, < div >$ e.t.c (in context of HTML).
- **Action Elements (A)**: UI elements which perform can post and get methods. Example: buttons, hyperlinks, submit buttons e.t.c.

- **Field Elements (F)**: UI elements that collect user input. Example: textbox, checkbox, radio buttons e.t.c.

Based on these 3 categorizations, I assume the following:

- User input is sent to server when a post method is called. It means every **F** is associated with a **A**. For instance the submit button comes after the search box. If there are multiple **F**'s while parsing through the tree it is associated with the same **A** within the branch. Every **F** has a corresponding **A** but **A** need not have a **F**.
- If there is **I**, it is linked with **A** either in its child nodes or child nodes to parent nodes of **I** to get the context.
- **I** can contain multiple **A**. For example, heading and list of links.
- **I** can have multiple **I**'s, each associated with '**A**'. For example, a webpage contains a heading, subheading and list of links for each subheading.

Based on the above assumptions, I parse through tree structure using a bottom-top approach and the tree structure can be broken into:
$I[I1, < A1, F1 >, I2, < A2, F2 >, I3, < A3, F3 >, ...]$
I1, I2, I3,...... contains all text information in a given node. $< An, Fn >$ is a set of A and its corresponding F pairs.

UI Selector

Recent studies Chung *et al.* (2022); Gu *et al.* (2023) propose training language models for discrimination rather than generation, as it enhances generalizability and sample efficiency for grounding tasks. I adopt this approach by transforming UI element selection into a multi-choice question answering problem. The language model

UI elements (U): For Calculator available UI elements are,
window : {
1, 2, 3, 4, 5, 6, 7, 8, 9, 0, =, +, , ×, ÷, ., +/, %, all clear, close button, zoom button, minimize button, main display: {0}
}
Menu bar : {
File: {Close, Close All, Save Tape As, Page Setup, Print Tape},
Edit: {Undo, Redo, Cut, Copy, Paste, Clear, Select All, Start Dictation, Emoji Symbols},
View: {Basic, Scientific, Programmer, Show Thousands Separators, RPN Mode, Decimal Places, Enter Full Screen}
Convert: {Recent Conversions, Area, Currency, Energy or Work, Length, Power, Pressure, Speed, Temperature, Time, Volume, Weights and Masses}
}
Example Query 7 : From the given list of available UI elements : U, select list of UI elements required for task "Convert 50 minutes to hour by dividing 50 with 60 in the Calculator" . The answer should be format [UI element 1, UI element 2, ...]
Response: [5, 0, ÷, 6, 0, =]

Figure 3.5: Illustrative example of UI selector selcting a list of UI elements.

is trained to select from a list of options instead of generating the complete target element. Once the UI elements are extracted, the UI elements required for a sub-task are selected by using a LLM as shown in Example Query 7 of Figure 3.5. If a UI element is not found, a feedback signal is sent to the conversational model, triggering a revision of the current sub-task.

Action Executor

The action executor is formulated with the purpose of executing a finite set of actions, including clicking, right-clicking, text entry and selection. An action executor is a python code that utilizes information about UI elements, such as their path, location and size, to execute an appropriate action based on the type of UI element. In order to enhance accuracy and reliability, the action executor is incorporated with a feedback mechanism to ascertain whether an action is performed or not. Once an action is performed, it generates action feedback for the next sub-task. Following the execution of all the sub-tasks, the action executor will send a message to the conversation model to generate an execution feedback and furnish results to users .

Chapter 4

NOVEL ASPECTS OF LUCI

Application-Centric Architecture

As seen in Figure. 3.1, the tool selector is the first component in my architecture. The conversational model generates sub-tasks based on the selected applications, which simplifies the problem statement and enables high-quality sub-task generation. The application-centric approach also **enable multi-application orchestration by separating the selection and orchestration tasks** between the Tool Selector and the Conversational Model.

Modular OS-Agnostic Agent

The Modular OS-Agnostic Agent represents a foundational aspect of LUCI's architecture. The decomposition of the entire architecture into independent and interchangeable components, each fulfilling specific functions within the framework, makes LUCI modular. Components such as the Tool Selector, Conversational Model, Task verifier and UI Selector are based on LLM's and are engineered to seamlessly operate across both native and web interfaces. Employing platform-independent techniques to traverse the UI hierarchy and extract relevant information, the UI Extractor within LUCI is designed to retrieve UI elements from desktop applications in a specific format, thus contributing to its operating system (OS) agnosticism. This design approach ensures LUCI's adaptability to diverse environments, facilitating its effectiveness across a range of operating systems and interface types.

Novel Tool Selection Mechanism

The Tool Selector embodies a novel tool selection mechanism aimed at identifying relevant tools for multi-application tasks. At its core, the Tool Selector incorporates a sophisticated mechanism driven by in-context learning, which enables it to discern the most relevant GUI tool from the available toolset. Unlike traditional selection methods, which may rely solely on predefined rules or heuristics, the Tool Selector dynamically adapts its decision-making process based on the context provided by the user's instructions and task-specific requirements. This adaptive approach allows the Tool Selector to consider various factors when identifying the optimal GUI tool for a given task. For instance, it may take into account the nature of the user's instructions, such as the specific actions or functionalities requested, as well as any contextual information provided, such as the current state of the application or the user's preferences. By leveraging this contextual awareness, the Tool Selector can make more informed decisions, ultimately leading to better tool selections and improved task performance.

Novel UI Parser

The contribution pertaining to the novel UI parser is exemplified by the UI Extractor component within LUCI. This component employs a novel UI parsing technique to extract structured IAF representations of both web and desktop interfaces. At its core, the UI Extractor employs a sophisticated UI parsing technique that allows it to traverse the complex hierarchy of UI elements present in web and desktop applications. By systematically categorizing these elements and extracting pertinent information such as type, location, size, and description, the UI Extractor generates structured representations that can be easily interpreted by the LLM. These IAF representations **sovle the limited context length issue** seen in previous methods.

Hierarchical Control Structure

This contribution is manifested in the hierarchical control structure implemented within LUCI, facilitating the control of tasks across multiple applications. At the heart of this hierarchical control structure lies the collaborative effort of key components such as the Conversational Model, Task Verifier, and Action Executor. The conversational model within LUCI employs in-context learning to understand user instructions and adapt its behavior accordingly continuously refining its understanding of tasks through feedback. Additionally, the Task Verifier filters redundant sub-tasks based on future tasks and past actions, minimizing unnecessary actions without human intervention. These features collectively allow LUCI to autonomously adapt to varying user needs and preferences without frequent human intervention while contributing to the overall effectiveness of the framework.

Chapter 5

EXPERIMENTS AND RESULTS

In this section, I evaluate the performance of LUCI and demonstrate that: (1) **LUCI outperforms previous approaches on executing complex tasks** (Section 5), **LUCI enables cross-application adaptability** (Section 5) and **LUCI can utilize multiple applications for executing complex tasks** (Section 5).

LUCI Outperforms Previous Approaches on Executing Complex Tasks

The performance of LUCI in executing complex tasks stands out prominently, especially when evaluated against other methodologies. my comprehensive examination involved testing LUCI across two benchmarks Miniwob++ Shi *et al.* (2017) and Mind2Web Deng *et al.* (2023b), to allow for fair comparisons with baselines. The MiniWoB++ task suite, designed to simulate real-world human-computer interactions Shi *et al.* (2017); Liu *et al.* (2018), poses simple tasks like click-checkboxes, and text-complete for computer agents. In my MiniWoB++ experiments, I performed experiments on two LLM's GPT-3.5-turbo and Phi-2, running 50 episodes to generate results for each task. In MiniWoB++ setup, I adopt RCI Kim *et al.* (2023) configurations with action space comprising of click-xpath, type, press, and click-options. The primary evaluation criterion is the success rate, reflecting the agent's effectiveness in completing the assigned task Kim *et al.* (2023). The success rate is determined by the proportion of successful episodes, wherein the agent receives a positive reward.

Mind2Web Deng *et al.* (2023b) is a realistic dataset with human demonstrations of open-domain tasks from diverse 137 real-world websites like Airbnb and Twitter, for assessing generalization across tasks, websites, and domains. For Mind2Web, I utilized GPT-3.5-turbo and Phi-2 with greedy decoding (i.e, temperature set to 0). Metrics include Operation F1 (Op. F1) for token-level F1 score for predicted operation comprised of action and input value, and Step SR for success rate per task step. This dataset is divided into three test sets: Cross-Task, Cross-Website, and Cross-Domain, evaluating generalizability over tasks from the same, similar and completely unseen domains, respectively. I include set of examples in prompts.

Further, to evaluate my work on day-to-day tasks I performed my experiments on 22 GUI applications (including both desktop and web applications) shown in Figure 5.2. To evaluate the performance on GUI applications, my key evaluation criterion is the success rate, reflecting the agent's effectiveness in completing the assigned task. Here, I've categorized three types of failure: tool selection, selection of unwanted UI elements and task failure. Additionally, an episode is deemed failed if the agent successfully carries out the created plan but is unable to complete the assignment and thus not rewarded. Most of the applications lack an appropriate dataset for comprehensive evaluation. To solve this problem, I employed an approach similar to Wang *et al.* (2023). I used a set of hand-written tasks serving as seed examples and then, utilize ChatGPT to generate more tasks. Unless explicitly stated otherwise, for these manually curated test sets, human evaluators assess and determine whether the task is considered to be accurately accomplished.

Performance on MiniWoB++ Task Suite

I performed comprehensive experiments to assess LUCI's performance in comparison to state-of-the-art (SOTA) approaches on MiniWoB++. For baseline comparisons using Behavior Cloning (BC) and Reinforcement Learning (RL), I employed CC-Net Humphreys *et al.* (2022b) and Pix2Act Shaw *et al.* (2023), which leverage large-scale BC and RL techniques. Regarding fine-tuning baselines, I evaluated against WebGUM Furuta *et al.* (2024b) and WebN-T5 Gur *et al.* (2023), two language models fine-tuned on a substantial number of demonstrations. In the realm of in-context learning (ICL) methods, my baselines comprised Synapse Zheng *et al.* (2024c), RCI Kim *et al.* (2023) and AdaPlanner Sun *et al.* (2023), both incorporating self-correction mechanisms. Additionally, I included human scores from Humphreys et al. Humphreys *et al.* (2022b) for supplementary benchmarking.

Figure 5.1 illustrates the average performance of different methods across Miniwob++ benchmark. LUCI, utilizing PHI2 and gpt-3.5-turbo, achieves human-level performance with mean success rates of 94% and 98.6%, respectively. Notably, LUCI

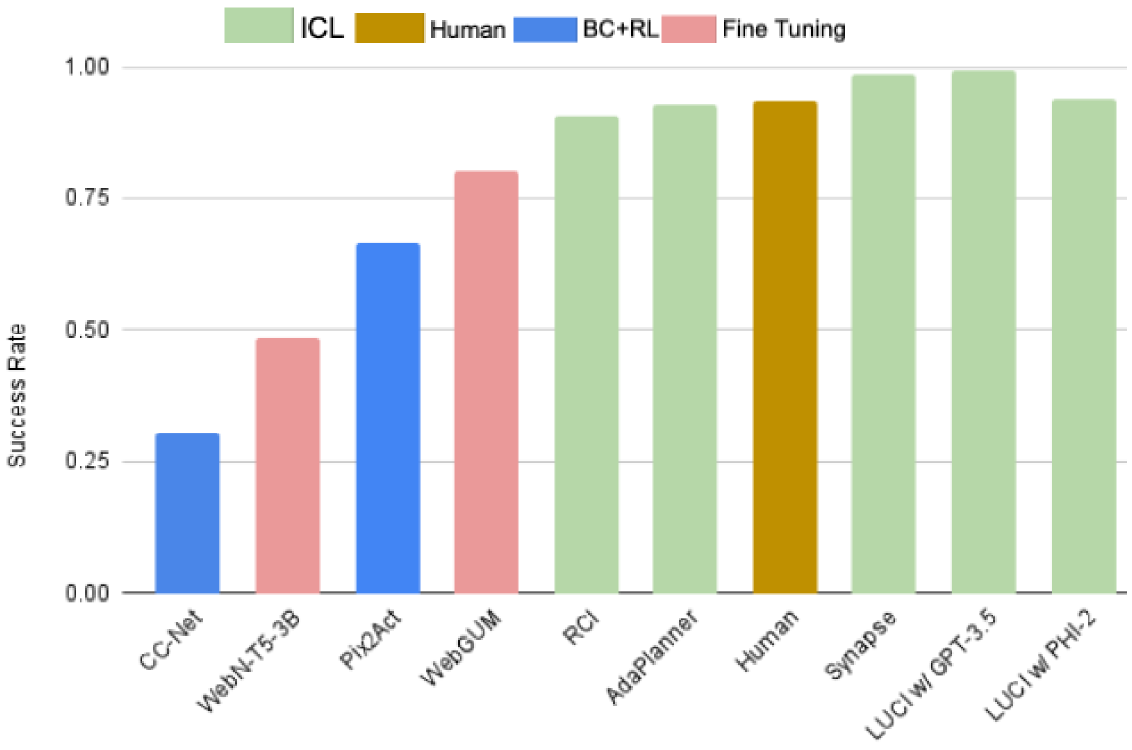

Figure 5.1: Average performance comparison with baselines in MiniWoB++ environment. LUCI w/ GPT-3.5 achieves state-of-the-art performance and LUCI w/ PHI-2 is the first model to achieve human level performance with LLM less than 3B parameters.

using gpt-3.5-turbo outperforms all baselines on MiniWoB++. LUCI surpasses previous ICL methods by addressing issues associated with context length, need for exemplar memory and human intervention for task adaptability. First, UI extractor in LUCI represents the user interface / HTML page in a compressed and structured way which solves tasks that previous methods cannot solve due to limited context length, such as book flight. Second, LUCI's hierarchical control structure and continuous user interaction enable dynamic task execution and adaptation without relying on exemplar memory, unlike Synapse Zheng *et al.* (2024c). **While Synapse and LUCI have around 99% success rate on the simpler Miniwob++ benchmark, LUCI shows nearly 3 times higher performance on Mind2WEB benchmark**. Third, conversational model within LUCI employs in-context learning to understand user instructions and adapt its behavior accordingly and continuously refining its understanding of tasks through feedback. Additionally, the Task Verifier filters redundant sub-tasks based on future tasks and past actions, minimizing unnecessary actions without human intervention. These features collectively allow LUCI to autonomously adapt to varying user needs and preferences without frequent human intervention. The instances of failure in LUCI are primarily attributed to the inherent challenges of tasks which cannot be planned ahead. For instance, tasks like tic-tac-toe, where the LUCI has to make dynamic decision-making at each turn, and the outcome of the game is contingent on the opponent's moves. Unlike other tasks which have deterministic or predictable outcomes, tic-tac-toe requires adaptability and the ability to react to the changing state of the game. LUCI cannot accurately plan ahead because it cannot foresee the opponent's moves beyond the current turn, making the traditional pre-planning approach less effective.

Table 5.1: Average performance of different methods on Mind2WEB Benchmark. LUCI w/ GPT-3.5 achieves state-of-the-art performance.

Baseline	Cross-Task		Cross-Website		Cross-Domain	
	Op. F1	Step SR	Op. F1	Step SR	Op. F1	Step SR
MINDACT	56.6	17.4	48.8	16.2	52.8	18.6
Synapse	-	30.6	-	29.1	-	26.4
WebGUM	75.9	64.9	75.3	62.5	77.7	66.7
GPT-4v	80.9	65.7	83.7	70	73.6	62.1
LUCI						
w/ GPT-3.5	93.8	86.7	96.3	89.1	91.7	84.2
w/ Phi2	82.3	72.8	84.9	77.3	79.4	69.1

Performance on Mind2Web

I showcase LUCI's applicability to real-world scenarios by testing it on Mind2Web Deng *et al.* (2023b). For baseline comparisons I used MindACT with GPT-3.5, WebGUM Furuta *et al.* (2024b), Gpt-4v Zheng *et al.* (2024b). The current SOTA in this benchmark is Gpt-4v Zheng *et al.* (2024b) with oracle grounding but requires human annotations. It requires LMM to generate action and then selects the UI element based on the action. In my experiments, I directly selects the UI element instead of generating action. LUCI with GPT-3.5-turbo achieves a Step success rate of 86.7 %, 89.1% and 84.2% across three test splits, respectively. As demonstrated in Table 5.1, my approach performs significantly better than other methods across three test splits over every metric. Notably, it achieves atleast 19% more in step success rate improvement over GPT-4(v) in all three settings using GPT-3.5. LUCI with Phi2 still performs admirably, demonstrating solid performance across various scenarios. It outperforms other models in most categories, showcasing the efficiency of LUCI with smaller LLMs in handling cross-task, cross-website, and cross-domain challenges.

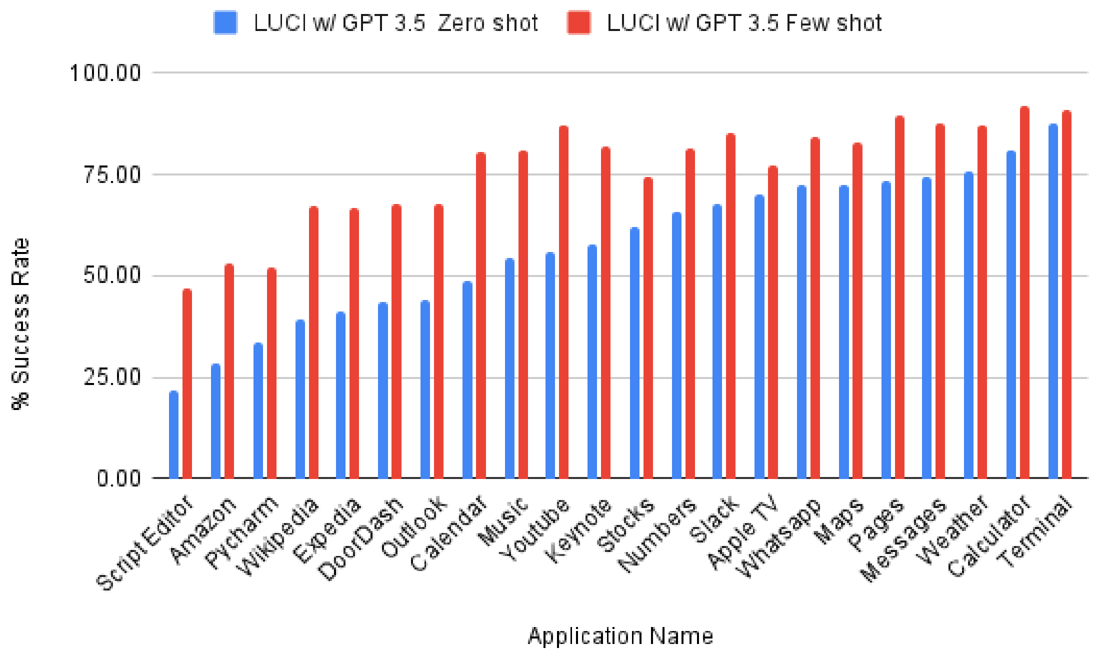

Figure 5.2: Average success rate of LUCI in using GUI Applications with GPT-3.5 under zero shot setting and Few Shot Setting

Performance of LUCI on GUI Applications

In this section I assess the effectiveness of my approach in empowering the model to autonomously leverage GUI applications, without the need for additional supervision. The results of my experiments, depicted in Figure 5.2, showcase the performance of LUCI when integrated with GPT-3.5 under both the zero-shot and few-shot in-context settings. Specifically, under the zero-shot setting, where the language model relies solely on its pre-existing knowledge to generate a solution outline, LUCI achieves an average success rate of 58%. In contrast, under the few-shot setting with limited context, the average success rate significantly increases to 76.5%, with over 60% of applications achieving a success rate of at least 80%. LUCI exhibits a comparatively lower performance in PyCharm, primarily attributed to the language model's limitations in generating accurate and effective code. LUCI demonstrates good performance on desktop applications even under the zero-shot setting when compared to web applications. However, a significant decrease is observed in the performance of LUCI with web applications under the few-shot setting. This disparity may stem from the language model's training data, which potentially contains information on how to navigate and interact with desktop applications but lacks comprehensive guidance on web applications. These findings highlight the LUCI's adaptability to scenarios where the model encounters unfamiliar domains with just few prompts.

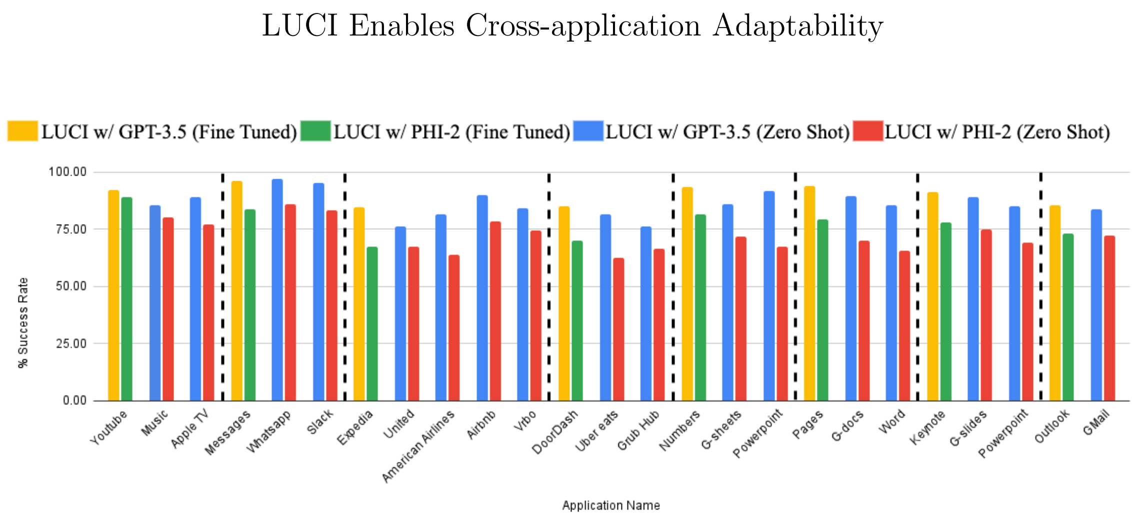

Figure 5.3: Cross application performance of LUCI with GPT-3.5 and PHI-2. LUCI fine-tuned on an application exhibits comparable performance on similar unseen applications. **LUCI can generalise to unseen environment.**

In this section, I closely examine the cross-application performance of language models with LUCI. Here I fine-tune language models on a single application and subsequently evaluating their success rate on analogous applications within same domains and task contexts. The models subjected to experimentation include GPT-3.5 Turbo and Phi-2. The objective of this investigation was to discern the adaptability of these agents when confronted with entirely new desktop or web applications, albeit within the familiarity of domains and task contexts they were originally fine-tuned for.

From Figure 5.3, it is observed that the models fine-tuned for a particular application exhibit a comparable success rate when tested on applications from the same domain. Quantitatively, the average deviation in performance is measured at 3.3 % for the GPT-3.5 Turbo setting and 4.5 % for the Phi-2 setting. This means that fine-tuning for a certain type of application helps the models do well on other applications in the same category.

LUCI Can Utilize Multiple Applications for Executing Complex Tasks

Another noteworthy aspect of LUCI is its ability to carry out tasks that require the integration of multiple applications. In this section, I evaluate LUCI's proficiency in seamlessly orchestrating various applications to efficiently execute multifaceted tasks, showcasing its potential for enhanced productivity and versatility in diverse user scenarios. To evaluate LUCI's capability to manage multiple applications, I created a set of hand-written tasks serving as seed examples and then, utilize ChatGPT to generate more tasks that requires the utilization of one or more GUI applications listed in Figure 5.2. Then, executed these tasks with number of GUI applications required to complete each task ranges from 1 to 6. In each case at least 21 tasks are evaluated and run 30 episodes to produce the results. My key evaluation criterion is the success rate discussed in Section 5, reflecting the agent's effectiveness in completing the assigned task. From Figure 5.4, delineates a trend wherein the success

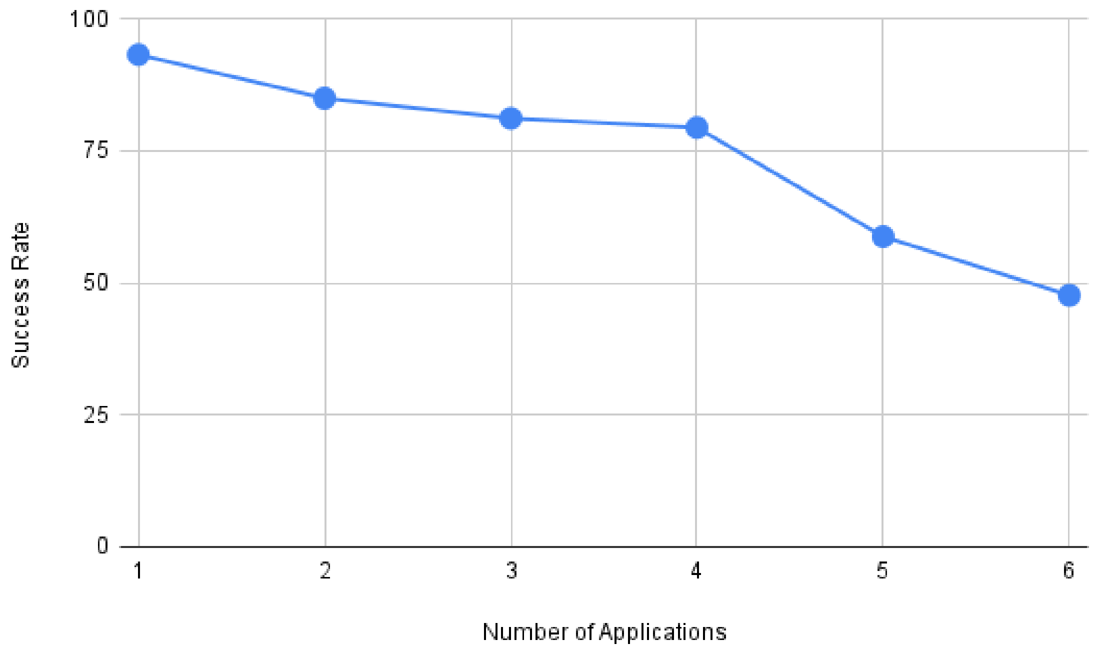

Figure 5.4: Average success rate of LUCI across tasks involving the use of multiple applications. The trend shows LUCI's ability to use at least four applications without losing efficacy.

rate exhibits a gradual decline from 93.17% for tasks involving a single application to 79.36% when four applications are concurrently utilized. This trend underscores LUCI's commendable performance in handling tasks comprising up to four applications. However, surpassing this threshold, the success rates sharply decrease to 58.73%, indicating a substantial challenge for LUCI in managing tasks necessitating the simultaneous usage of more than five applications. These findings underscore the diminishing efficacy of LUCI with an increasing number of applications, implying complexities in its multitasking capabilities beyond a certain threshold.

During my experiments, I note that the ordering of applications can have significant impact on the success rate of entire task. While executing tasks involving multiple applications, complex applications such as Keynote, are called in later stages leads to a lower success rate. I can attribute this effect to long term attention limitations in LLMs. This implies a LUCI with advanced LLM (such as GPT-4) can alleivate these issues.

Chapter 6

FUTURE WORK

LUCI is designed to simplify the creation and evaluation of versatile agents optimized for GUI tools. These agents show great potential in improving the accessibility and usability of GUI tools, especially for those who are unfamiliar with information technology or have impairments that may make it difficult to navigate complex tools or applications. Despite its potential benefits, there remain significant concerns and limits regarding present data gathering approaches, system design, and the necessary safety precautions for deployment in real-world circumstances.

Representation in Data: The data and methodology have undergone evaluation for English instructions and user interfaces containing English text. In future, I would expand to different languages.

Use of Multimodal Information: LUCI, focuses on modeling the GUI environment into textual context from underlying hierarchy, neglecting other information such as images, videos e.t.c. This makes LUCI vulnerable to performs actions based on information other than text. Leveraging this multimodal information holds promise for enhancing model performance.

Tolerance to Noise: In LUCI, a solution outline is generate ahead of execution based on previous knowledge. Deviations of desktop or web application from the original user interfaces, often triggered by Pop-ups and Ads, result in errors as LUCI struggles to adjust to unexpected scenarios.

Safety Concerns: The development of general-purpose action agents holds the potential to enhance efficiency and user experiences but requires careful consideration of safety concerns. Key issues include managing sensitive actions, privacy-related activities, and the risk of breaching security measures. Amid these challenges, action agents pose a significant risk of breaching security involving authentication and authorization processes, including CAPTCHA, and may be exploited for malicious activities. A comprehensive approach is needed for responsible deployment, urging proactive cybersecurity research to develop preemptive protective measures.

Chapter 7

CONCLUSION

In this work, I introduced LLM assisted User Control Interface (LUCI), a computer agent that leverages the reasoning capabilities of LLMs, such as GPT-3.5 and PHI-2, to interact and control wide range of desktop and web applications to execute repetitive actions and solve complex tasks. LUCI addresses context-length issues as seen in previous methods by using compressed semantic representations for UI elements across both native and web interfaces. This extends the capabilities of previous single platform approaches. Additionally, LUCI leverages a hierarchical structure enabling multi-application control. LUCI accomplishes all this while maintaining similar or up to 20% better performance on the benchmarks like MiniWoB++, Mind2Web.

REFERENCES

Ahn, M., A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan and A. Zeng, "Do as i can, not as i say: Grounding language in robotic affordances", (2022).

Chung, H. W., L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le and J. Wei, "Scaling instruction-finetuned language models", (2022).

Dasgupta, I., C. Kaeser-Chen, K. Marino, A. Ahuja, S. Babayan, F. Hill and R. Fergus, "Collaborating with language models for embodied reasoning", (2023a).

Dasgupta, I., C. Kaeser-Chen, K. Marino, A. Ahuja, S. Babayan, F. Hill and R. Fergus, "Collaborating with language models for embodied reasoning", (2023b).

Deng, X., Y. Gu, B. Zheng, S. Chen, S. Stevens, B. Wang, H. Sun and Y. Su, "Mind2web: Towards a generalist agent for the web", (2023a).

Deng, X., Y. Gu, B. Zheng, S. Chen, S. Stevens, B. Wang, H. Sun and Y. Su, "Mind2web: Towards a generalist agent for the web", (2023b).

Desai, S., T. Sharma and P. Saha, "Using chatgpt in hci research—a trioethnography", in "Proceedings of the 5th International Conference on Conversational User Interfaces", CUI '23 (ACM, 2023), URL `http://dx.doi.org/10.1145/3571884.3603755`.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale", (2021).

Furuta, H., K.-H. Lee, O. Nachum, Y. Matsuo, A. Faust, S. S. Gu and I. Gur, "Multimodal web navigation with instruction-finetuned foundation models", (2024a).

Furuta, H., K.-H. Lee, O. Nachum, Y. Matsuo, A. Faust, S. S. Gu and I. Gur, "Multimodal web navigation with instruction-finetuned foundation models", (2024b).

Gao, L., A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan and G. Neubig, "Pal: Program-aided language models", (2023).

Glaese, A., N. McAleese, M. Trebacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker, L. Campbell-Gillingham, J. Uesato, P.-S. Huang, R. Comanescu, F. Yang, A. See, S. Dathathri, R. Greig, C. Chen, D. Fritz, J. S. Elias, R. Green, S. Mokrá, N. Fernando, B. Wu, R. Foley, S. Young, I. Gabriel, W. Isaac, J. Mellor, D. Hassabis, K. Kavukcuoglu, L. A. Hendricks and G. Irving, "Improving alignment of dialogue agents via targeted human judgements", (2022).

Gu, Y., X. Deng and Y. Su, "Don't generate, discriminate: A proposal for grounding language models to real-world environments", (2023).

Gur, I., N. Jaques, Y. Miao, J. Choi, M. Tiwari, H. Lee and A. Faust, "Environment generation for zero-shot compositional reinforcement learning", (2022).

Gur, I., O. Nachum, Y. Miao, M. Safdari, A. Huang, A. Chowdhery, S. Narang, N. Fiedel and A. Faust, "Understanding html with large language models", (2023).

Gur, I., U. Rueckert, A. Faust and D. Hakkani-Tur, "Learning to navigate the web", (2018).

Hoffmann, J., S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals and L. Sifre, "Training compute-optimal large language models", (2022).

Hong, W., W. Wang, Q. Lv, J. Xu, W. Yu, J. Ji, Y. Wang, Z. Wang, Y. Zhang, J. Li, B. Xu, Y. Dong, M. Ding and J. Tang, "Cogagent: A visual language model for gui agents", (2023).

Huang, W., P. Abbeel, D. Pathak and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents", in "Proceedings of the 39th International Conference on Machine Learning", edited by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu and S. Sabato, vol. 162 of *Proceedings of Machine Learning Research*, pp. 9118–9147 (PMLR, 2022a), URL `https://proceedings.mlr.press/v162/huang22a.html`.

Huang, W., F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, P. Sermanet, N. Brown, T. Jackson, L. Luu, S. Levine, K. Hausman and B. Ichter, "Inner monologue: Embodied reasoning through planning with language models", (2022b).

Humphreys, P. C., D. Raposo, T. Pohlen, G. Thornton, R. Chhaparia, A. Muldal, J. Abramson, P. Georgiev, A. Goldin, A. Santoro and T. Lillicrap, "A data-driven approach for learning to control computers", (2022a).

Humphreys, P. C., D. Raposo, T. Pohlen, G. Thornton, R. Chhaparia, A. Muldal, J. Abramson, P. Georgiev, A. Goldin, A. Santoro and T. Lillicrap, "A data-driven approach for learning to control computers", (2022b).

Isyanto, H., A. S. Arifin and M. Suryanegara, “Design and implementation of iot-based smart home voice commands for disabled people using google assistant”, in “2020 International Conference on Smart Technology and Applications (ICoSTA)”, pp. 1–6 (2020).

Jia, S., J. Kiros and J. Ba, “Dom-q-net: Grounded rl on structured language”, (2019).

Kim, G., P. Baldi and S. McAleer, “Language models can solve computer tasks”, (2023).

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa, “Large language models are zero-shot reasoners”, (2023).

Lazaridou, A., E. Gribovskaya, W. Stokowiec and N. Grigorev, “Internet-augmented language models through few-shot prompting for open-domain question answering”, (2022).

Li, T. J.-J., A. Azaria and B. A. Myers, “Sugilite: Creating multimodal smartphone automation by demonstration”, in “Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems”, CHI ’17, p. 6038–6049 (Association for Computing Machinery, New York, NY, USA, 2017), URL `https://doi.org/10.1145/3025453.3025483`.

Liang, Y., C. Wu, T. Song, W. Wu, Y. Xia, Y. Liu, Y. Ou, S. Lu, L. Ji, S. Mao, Y. Wang, L. Shou, M. Gong and N. Duan, “Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis”, (2023).

Liu, E. Z., K. Guu, P. Pasupat, T. Shi and P. Liang, “Reinforcement learning on web interfaces using workflow-guided exploration”, (2018).

Menick, J., M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving and N. McAleese, “Teaching language models to support answers with verified quotes”, (2022).

Mialon, G., R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun and T. Scialom, “Augmented language models: a survey”, (2023).

Nakano, R., J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess and J. Schulman, “Webgpt: Browser-assisted question-answering with human feedback”, (2022).

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike and R. Lowe, “Training language models to follow instructions with human feedback”, (2022).

Paranjape, B., S. Lundberg, S. Singh, H. Hajishirzi, L. Zettlemoyer and M. T. Ribeiro, “Art: Automatic multi-step reasoning and tool-use for large language models”, (2023).

Pasupat, P., T.-S. Jiang, E. Liu, K. Guu and P. Liang, "Mapping natural language commands to web elements", in "Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing", edited by E. Riloff, D. Chiang, J. Hockenmaier and J. Tsujii, pp. 4970–4976 (Association for Computational Linguistics, Brussels, Belgium, 2018), URL `https://aclanthology.org/D18-1540`.

Schick, T., J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda and T. Scialom, "Toolformer: Language models can teach themselves to use tools", (2023).

Shaw, P., M. Joshi, J. Cohan, J. Berant, P. Pasupat, H. Hu, U. Khandelwal, K. Lee and K. Toutanova, "From pixels to ui actions: Learning to follow instructions via graphical user interfaces", (2023).

Shi, T., A. Karpathy, L. Fan, J. Hernandez and P. Liang, "World of bits: An open-domain platform for web-based agents", in "Proceedings of the 34th International Conference on Machine Learning", edited by D. Precup and Y. W. Teh, vol. 70 of *Proceedings of Machine Learning Research*, pp. 3135–3144 (PMLR, 2017), URL `https://proceedings.mlr.press/v70/shi17a.html`.

Sun, H., Y. Zhuang, L. Kong, B. Dai and C. Zhang, "Adaplanner: Adaptive planning from feedback with language models", (2023).

Taylor, R., M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez and R. Stojnic, "Galactica: A large language model for science", (2022).

Thoppilan, R., D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi and Q. Le, "Lamda: Language models for dialog applications", (2022).

Tulshan, A. S. and S. N. Dhage, "Survey on virtual assistant: Google assistant, siri, cortana, alexa", in "Advances in Signal Processing and Intelligent Recognition Systems", edited by S. M. Thampi, O. Marques, S. Krishnan, K.-C. Li, D. Ciuonzo and M. H. Kolekar, pp. 190–201 (Springer Singapore, Singapore, 2019).

Vemprala, S., R. Bonatti, A. Bucker and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities", (2023).

Wang, Y., Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions", (2023).

Wei, J., M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai and Q. V. Le, "Finetuned language models are zero-shot learners", (2022).

Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models”, (2023).

Yang, R., L. Song, Y. Li, S. Zhao, Y. Ge, X. Li and Y. Shan, “Gpt4tools: Teaching large language model to use tools via self-instruction”, (2023).

Yao, S., H. Chen, J. Yang and K. Narasimhan, “Webshop: Towards scalable real-world web interaction with grounded language agents”, (2023).

Ye, J., X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang and X. Huang, “A comprehensive capability analysis of gpt-3 and gpt-3.5 series models”, (2023).

Zeng, A., M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke and P. Florence, “Socratic models: Composing zero-shot multimodal reasoning with language”, (2022).

Zheng, B., B. Gou, J. Kil, H. Sun and Y. Su, “Gpt-4v(ision) is a generalist web agent, if grounded”, (2024a).

Zheng, B., B. Gou, J. Kil, H. Sun and Y. Su, “Gpt-4v(ision) is a generalist web agent, if grounded”, (2024b).

Zheng, L., R. Wang, X. Wang and B. An, “Synapse: Trajectory-as-exemplar prompting with memory for computer control”, (2024c).

Zhou, S., F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried, U. Alon and G. Neubig, “Webarena: A realistic web environment for building autonomous agents”, (2023).