UIST PITTSBURGH | 2024

LlamaTouch: A Faithful and Scalable Testbed for Mobile UI Task Automation

Li Zhang, Shihe Wang, Xianqing Jia, Zhihan Zheng, Yunhe Yan, Longxi Gao, Yuanchun Li[#], Mengwei Xu

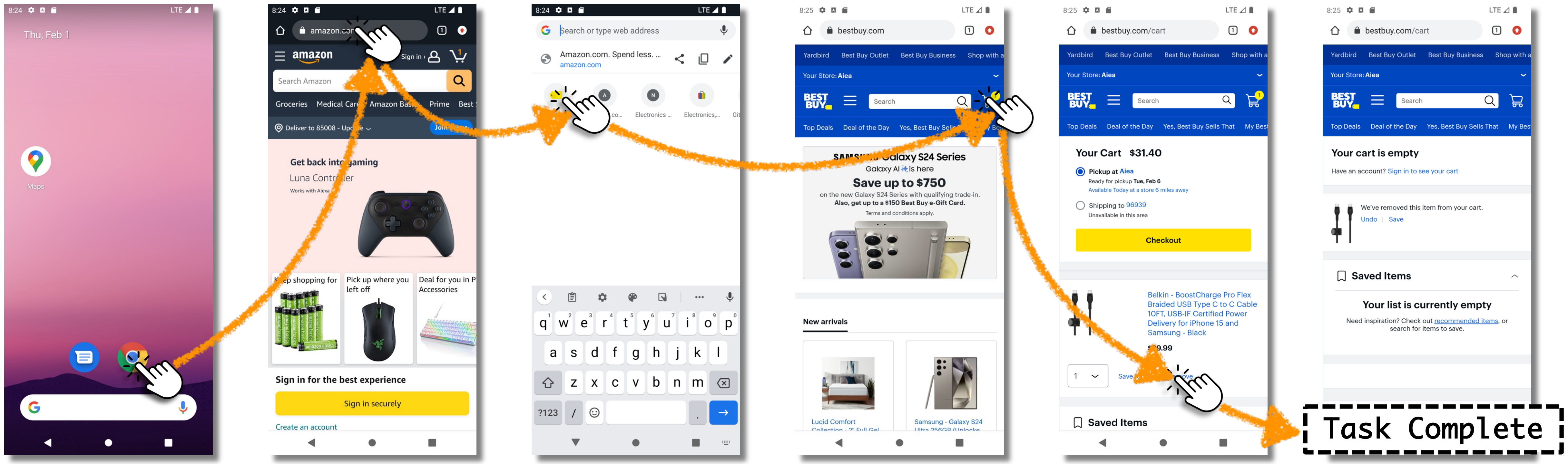
Beijing University of Posts and Telecommunications (BUPT)
[#]Institute for AI Industry Research (AIR), Tsinghua University

Mobile UI Task Automation

: Hey Mobile Agent!

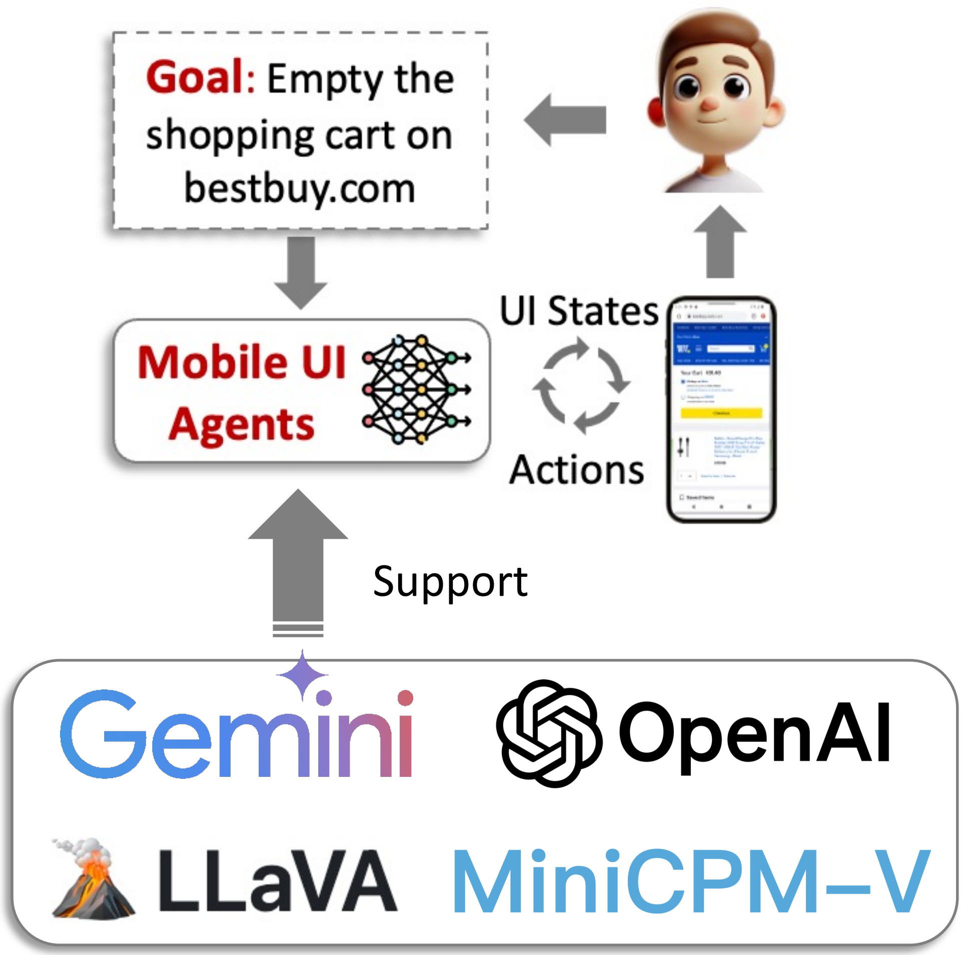
Empty the shopping cart on BestBuy!

: Sure!



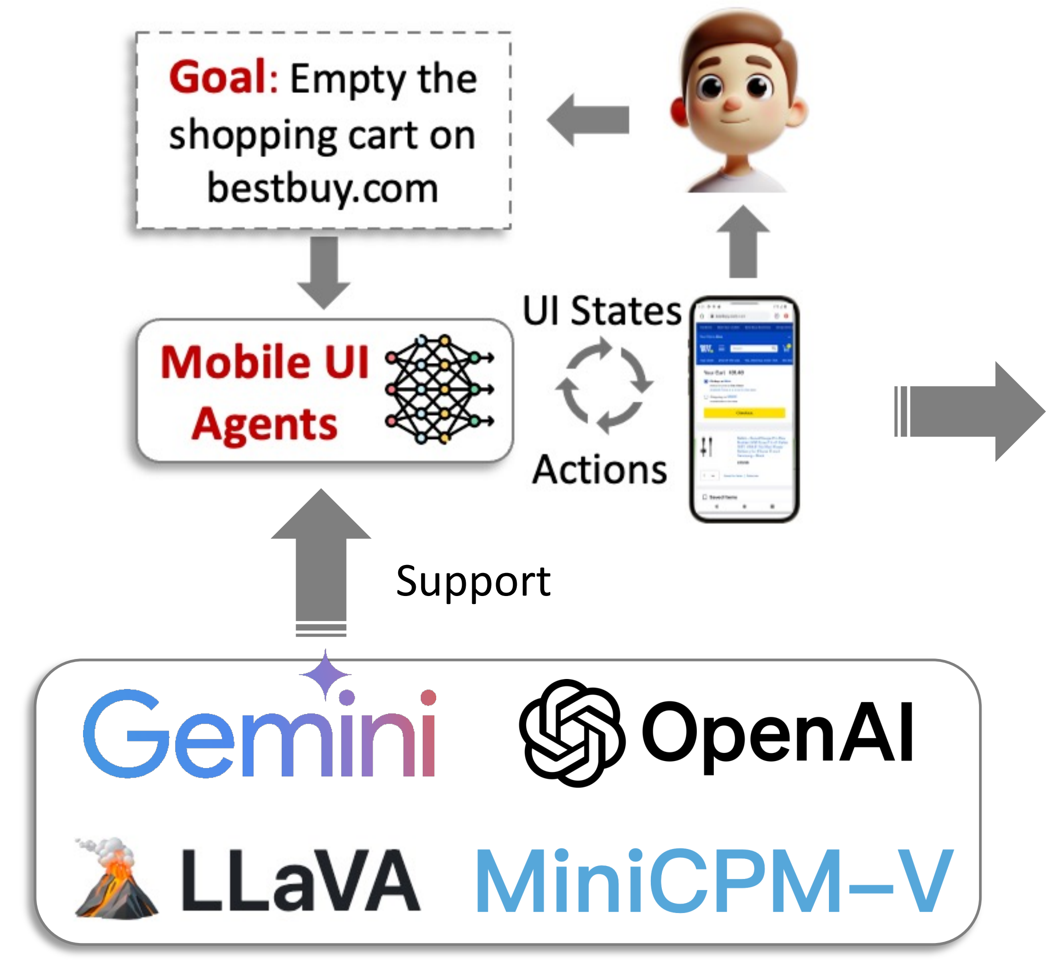
LLM/MLLM-powered Mobile UI Agents

LLM/MLLMs for Mobile UI Perception and Action Prediction

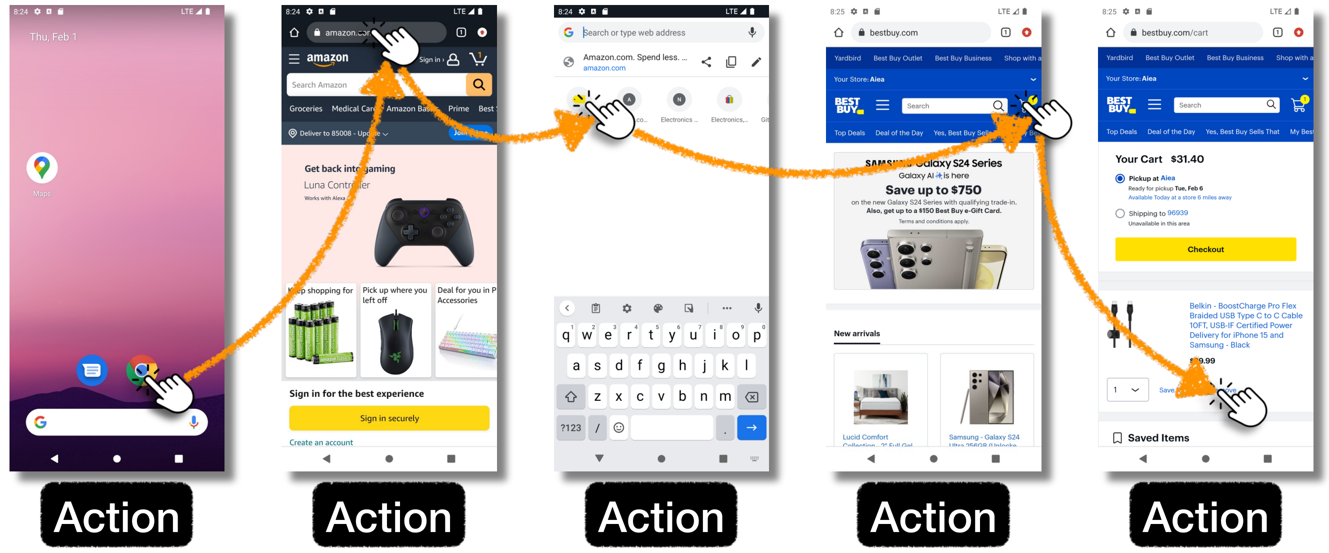


LLM/MLLM-powered Mobile UI Agents

LLM/MLLMs for Mobile UI Perception and Action Prediction



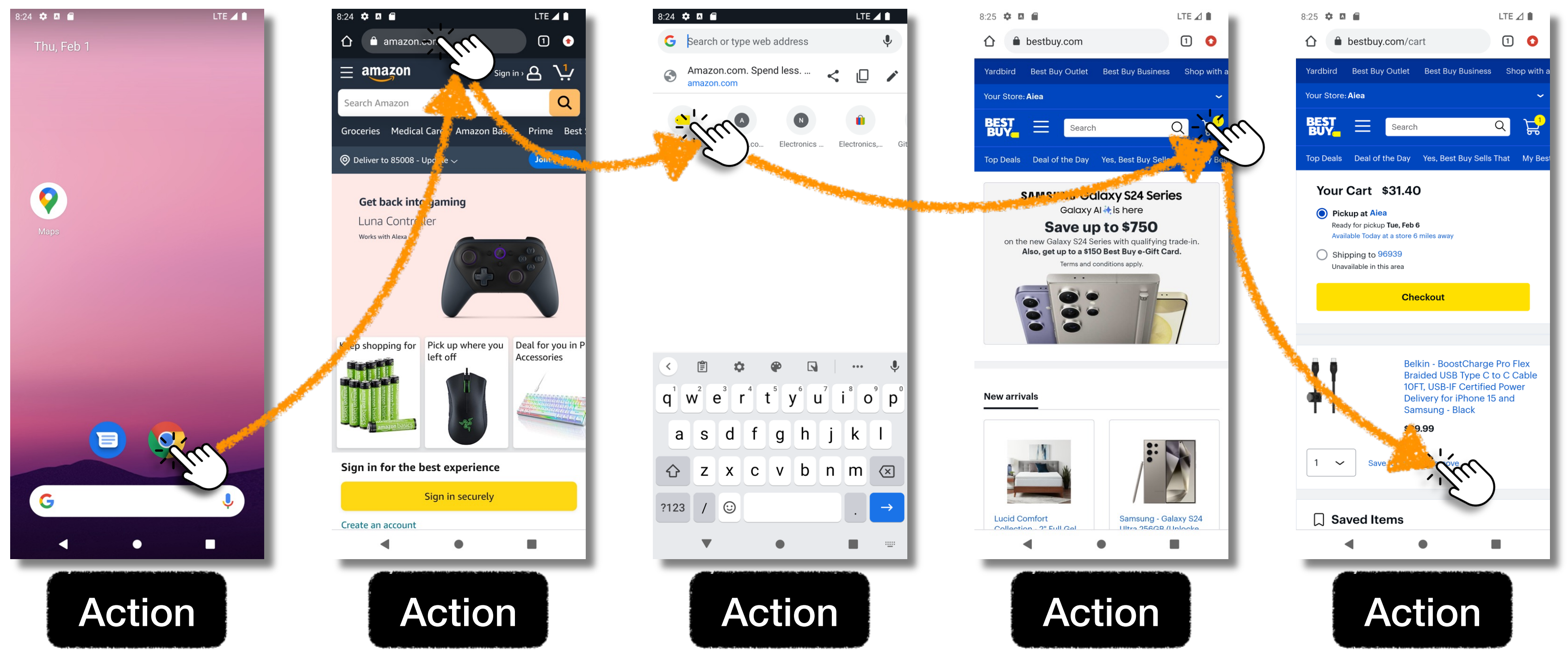
Task Execution Traces



Can these agents accurately complete user requests?

Evaluating Mobile UI Agents

Task Execution Traces

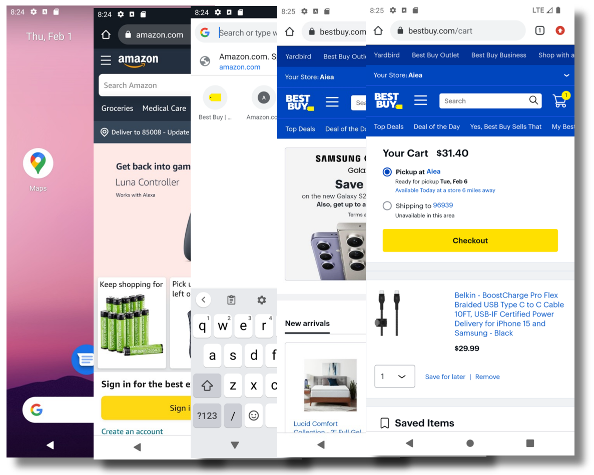


Approach #1:
Human validation

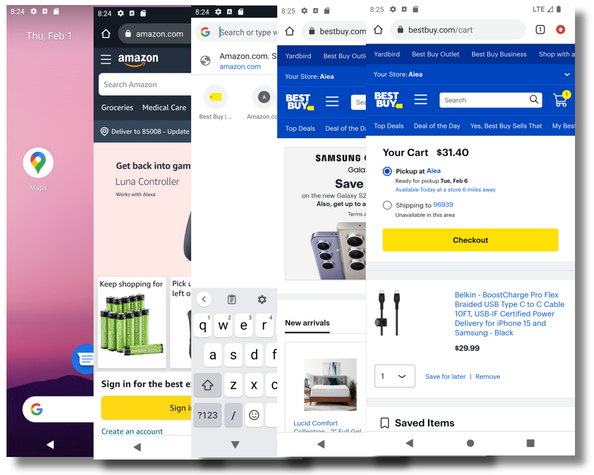
- ✓ High accuracy
- Low scalability
- High cost

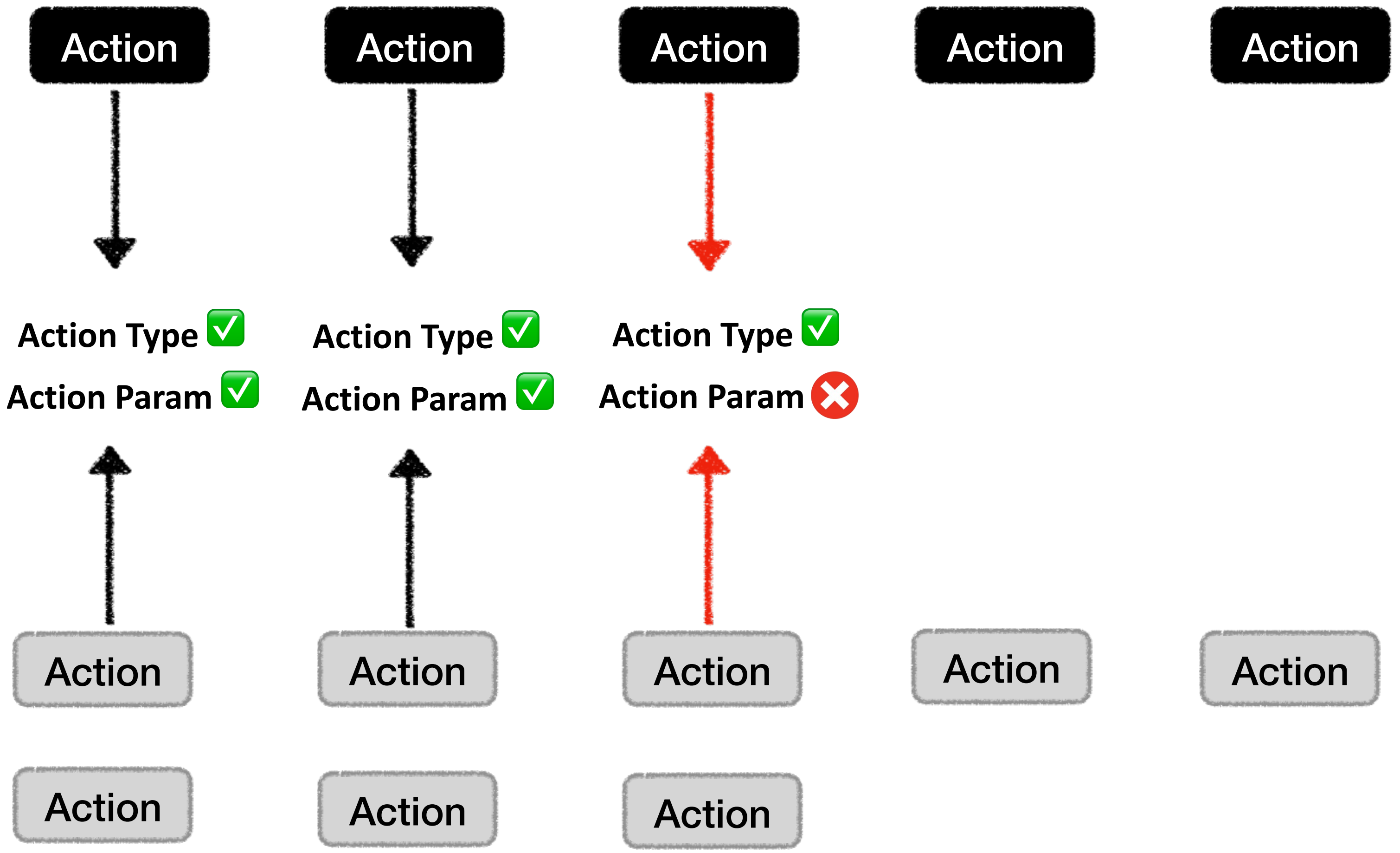
Evaluating Mobile UI Agents

Task Execution Traces



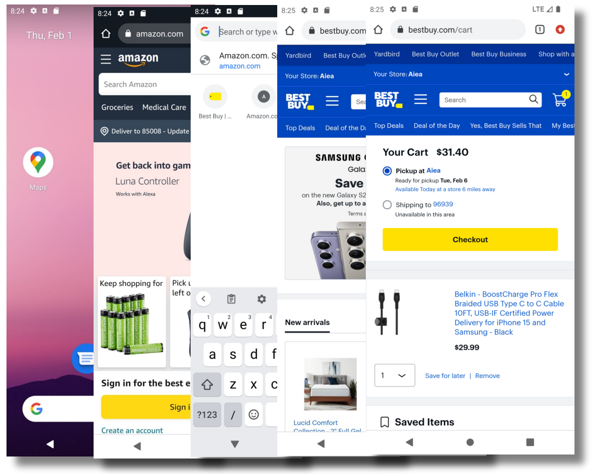
Ground-truth UI Traces in Static Datasets



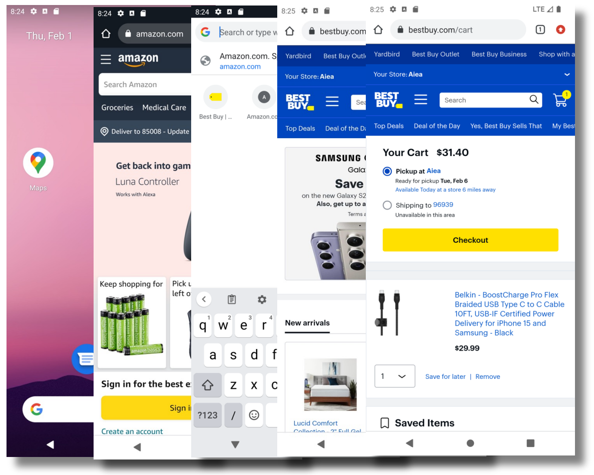


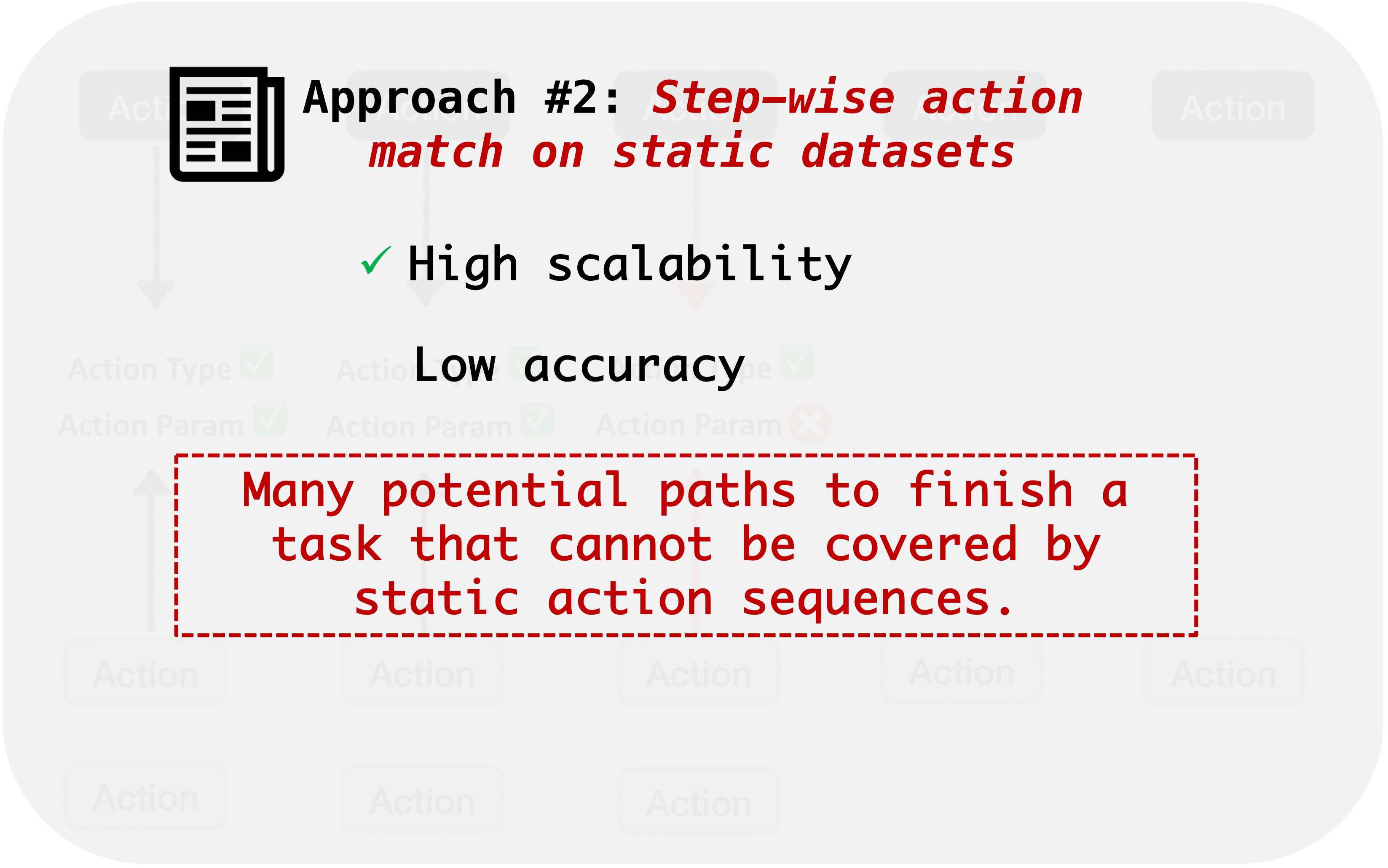
Evaluating Mobile UI Agents

Task Execution Traces



Ground-truth UI Traces in Static Datasets





Approach #2: ***Step-wise action match on static datasets***

✓ High scalability

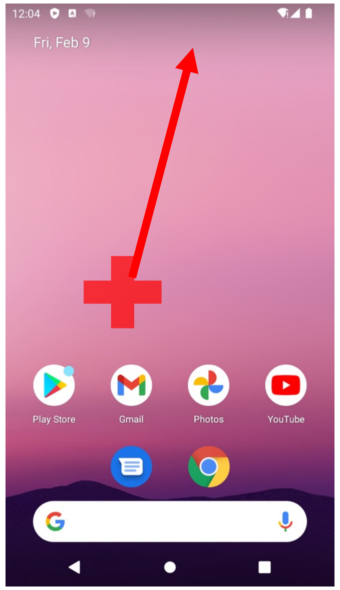
Low accuracy

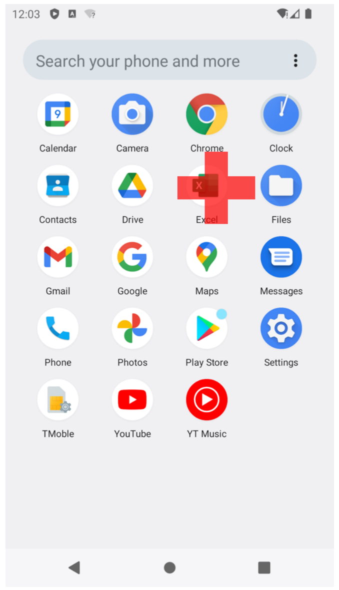
Many potential paths to finish a task that cannot be covered by static action sequences.

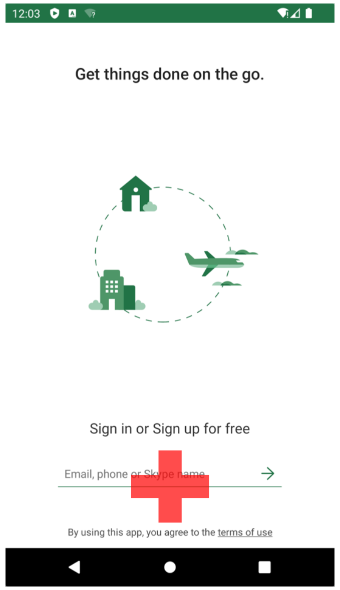
Observation

- Target: <u>Scalable</u> (as evaluated on static datasets) and <u>faithful</u> (as evaluated by humans) evaluation
- Observation: UI automation tasks transfer app states represented on the screen.

Task: Open app "Microsoft Excel" (install if not already installed), go to the login page



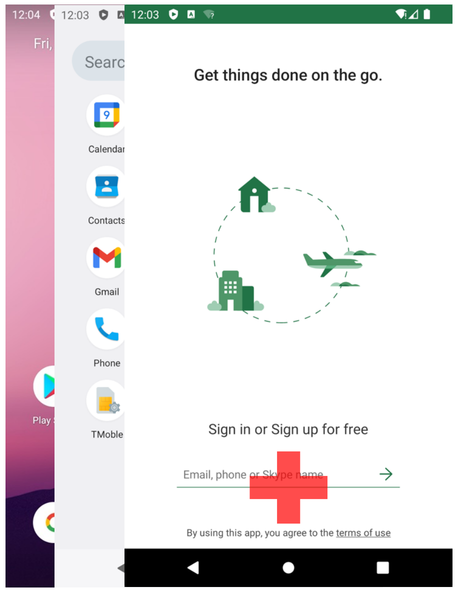


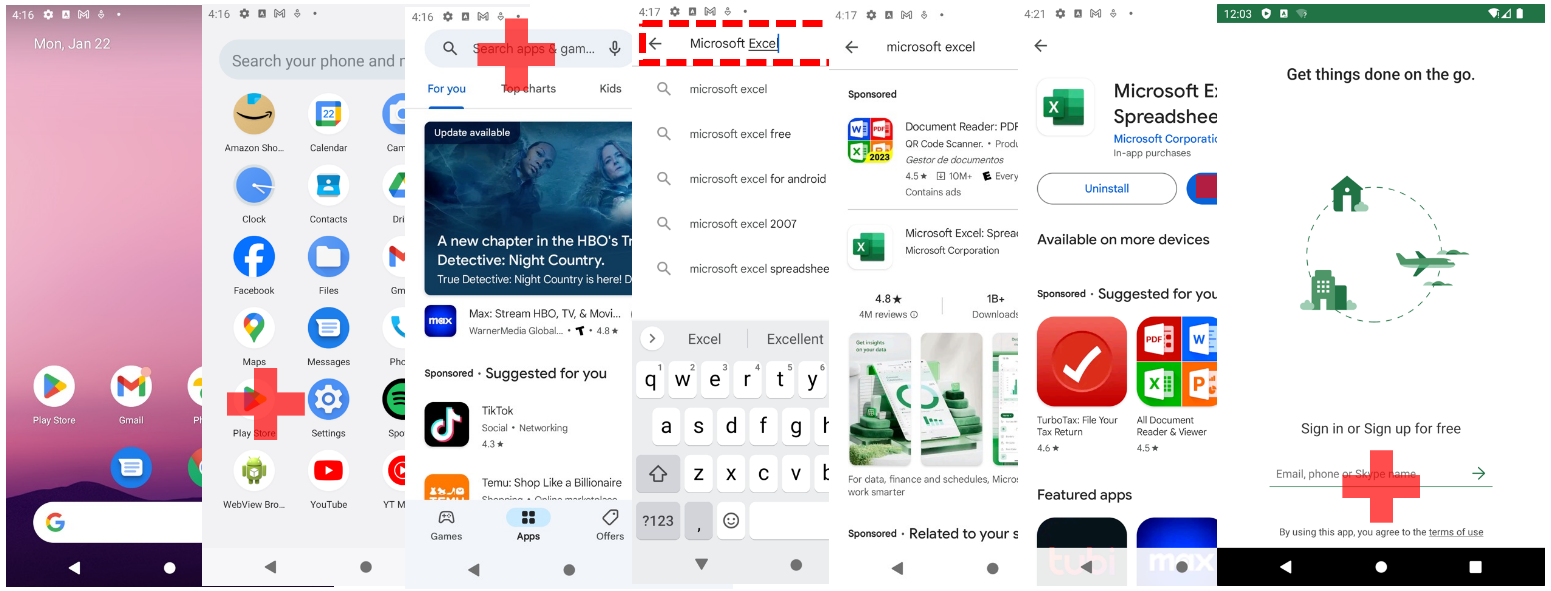

12:03
Get things done on the go.
Sign in or Sign up for free
Email, phone or Skype name
By using this app, you agree to the terms of use

Observation

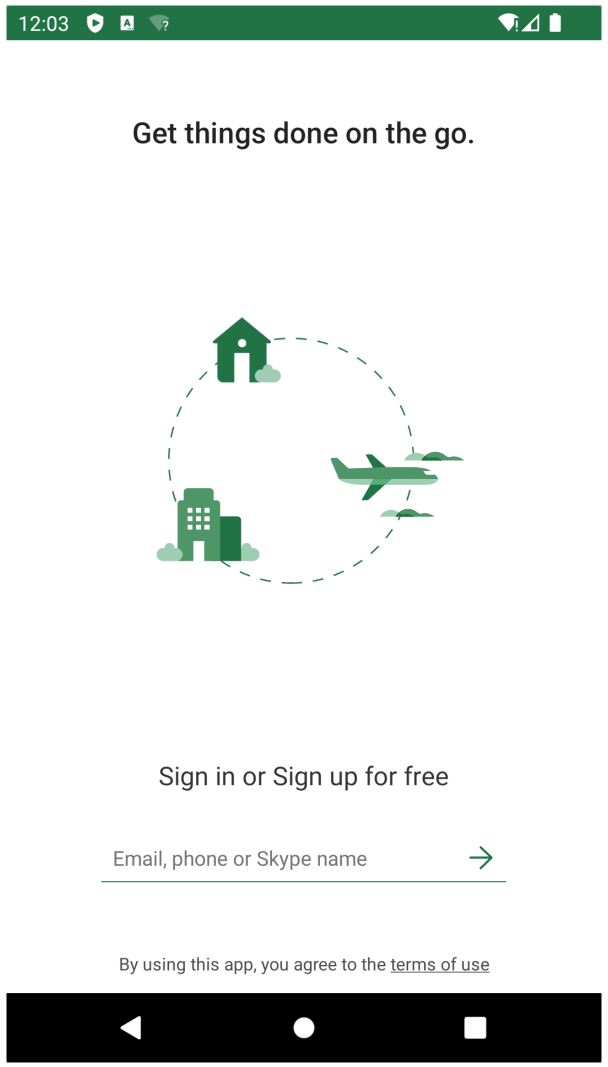
- ❑ Target: <u>Scalable</u> (as evaluated on static datasets) and <u>faithful</u> (as evaluated by humans) evaluation
- ❑ Observation: UI automation tasks transfer app states represented on the screen.

Task: Open app "Microsoft Excel" (install if not already installed), go to the login page


Get things done on the go.
Sign in or Sign up for free
Email, phone or Skype name
By using this app, you agree to the terms of use


Play Store
Search apps & games
Microsoft Excel
microsoft excel
Microsoft Excel: Spreadsheet
Microsoft Corporation
Uninstall
Get things done on the go.
Sign in or Sign up for free
Email, phone or Skype name

Observation


12:03

Get things done on the go.

Sign in or Sign up for free

Email, phone or Skype name

By using this app, you agree to the terms of use

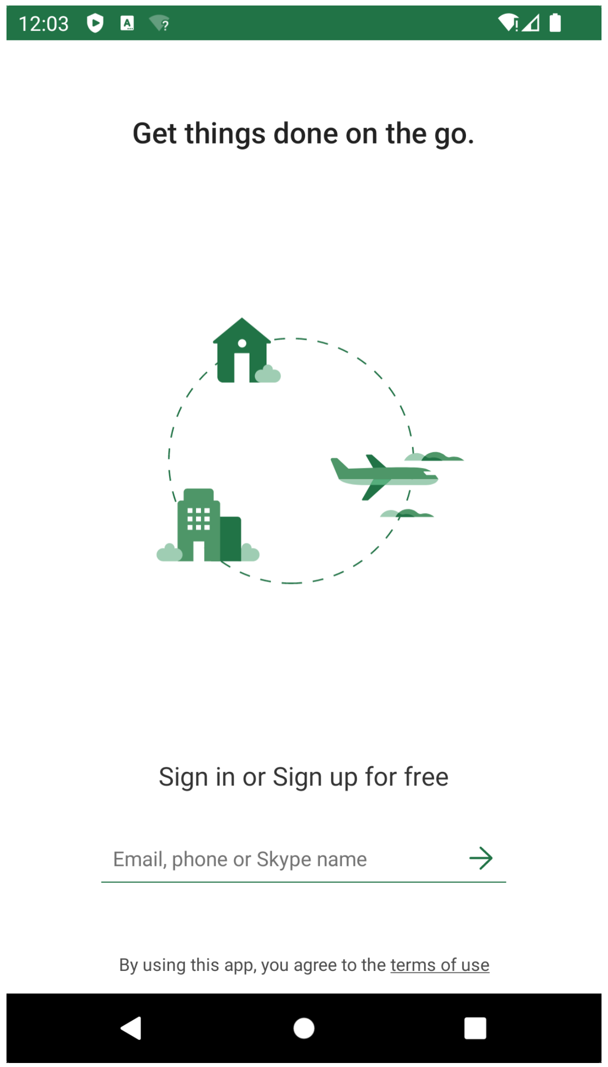
Different task execution paths

$\downarrow$

The same app state

$\downarrow$

The login page of Microsoft Excel


12:03

Get things done on the go.

Sign in or Sign up for free

Email, phone or Skype name

By using this app, you agree to the terms of use

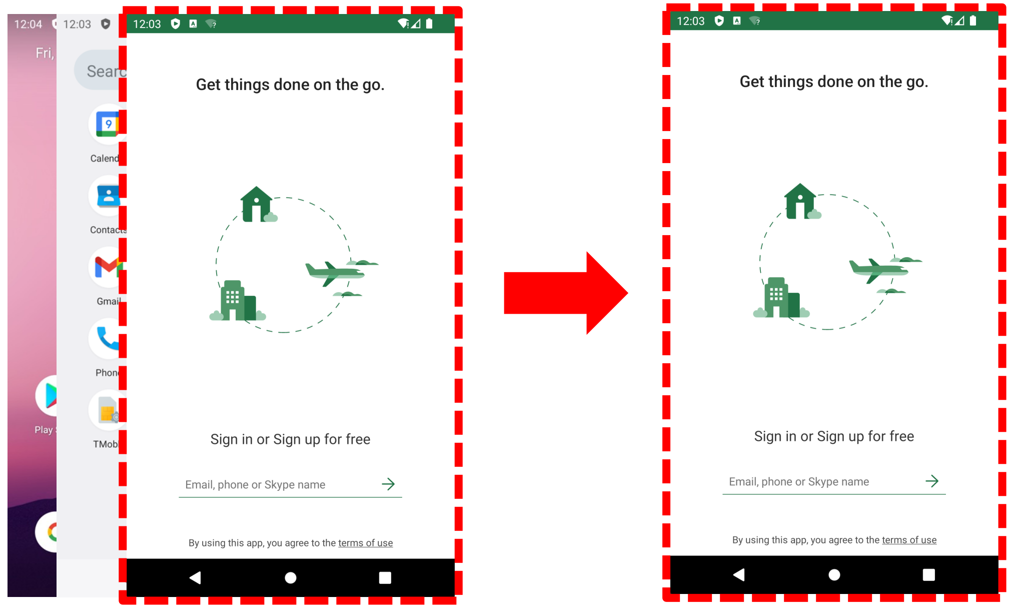
Our Approach

- **Our approach: Check app states during and after task execution, rather than comparing concrate action sequences.**

For a given task, how to annotate app states to represent task completion?

Task Completion Annotation

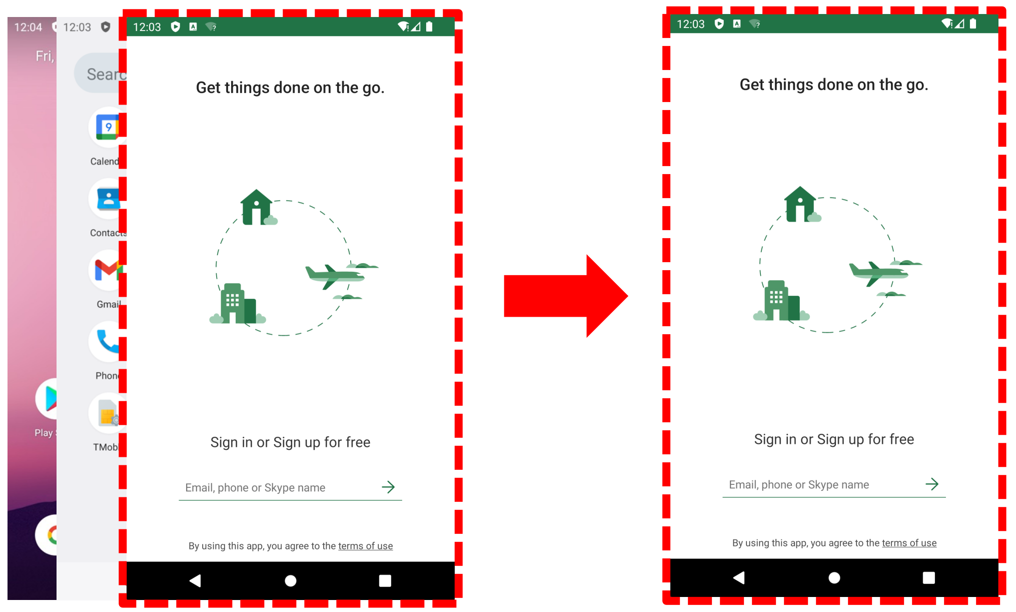
- ❑ **Annotate and match at the whole-screen level**


12:04
12:03
Fri,
Search
Calendar
Contacts
Gmail
Phone
Play
TMob
Get things done on the go.
Sign in or Sign up for free
Email, phone or Skype name
By using this app, you agree to the terms of use

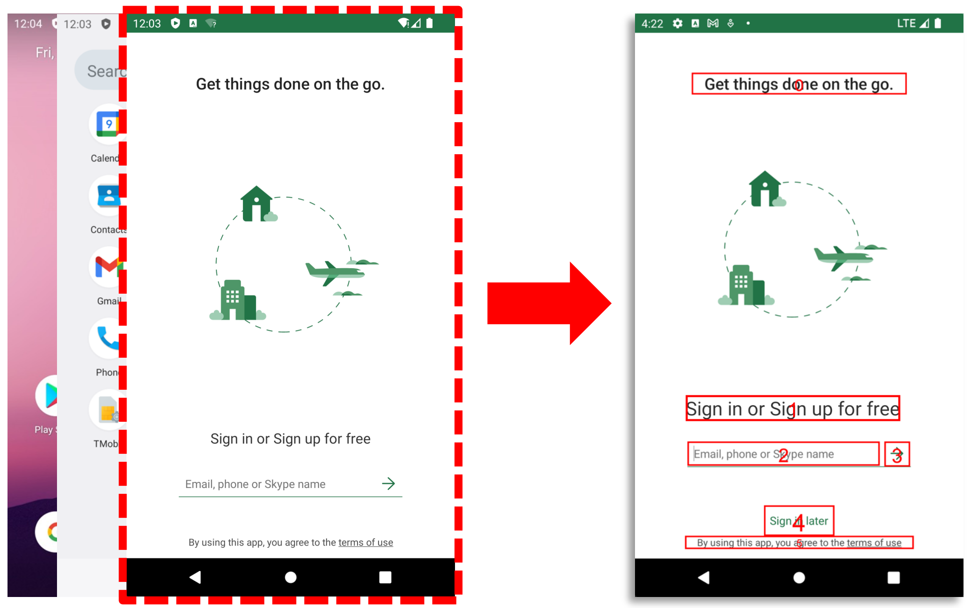
Unable to handle dynamic screen resolutions, dynamic screen contents (e.g., ads)

Task Completion Annotation

- ❑ Annotate and match at the whole-screen level
- ❑ What we want: Fine-grained app state annotation



Get things done on the go.
Sign in or Sign up for free
Email, phone or Skype name
By using this app, you agree to the terms of use

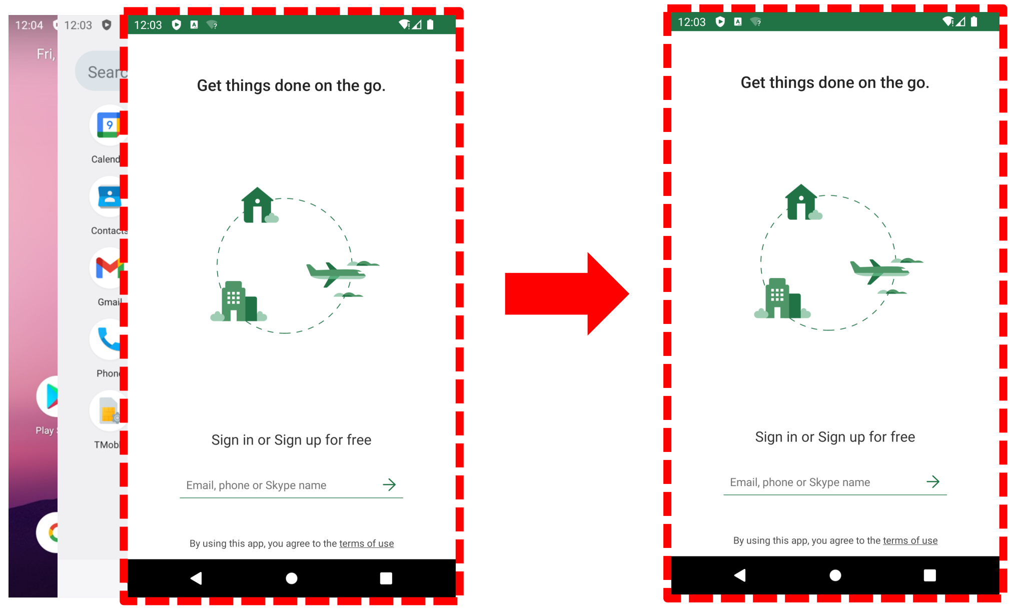


Get things done on the go.
Sign in or Sign up for free
Email, phone or Skype name
Sign in later
By using this app, you agree to the terms of use
0
1
2
3
4
5

Unable to handle dynamic screen resolutions, dynamic screen contents (e.g., ads)

Task Completion Annotation

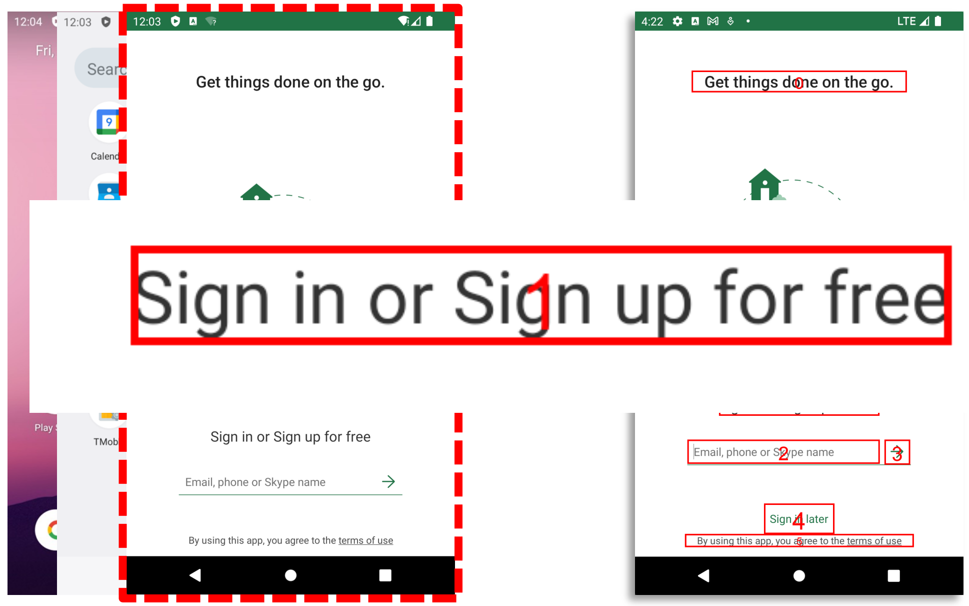
- ❑ Annotate and match at the whole-screen level



Get things done on the go.
Sign in or Sign up for free
Email, phone or Skype name
By using this app, you agree to the terms of use

Unable to handle dynamic screen resolutions, dynamic screen contents (e.g., ads)

- ❑ What we want: Fine-grained app state annotation

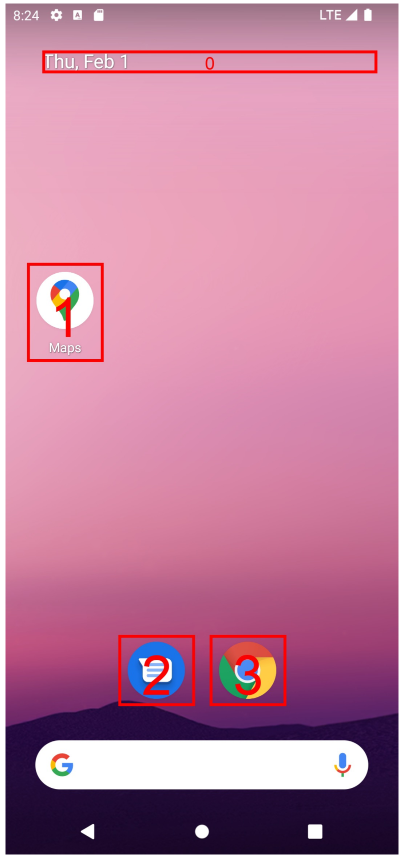


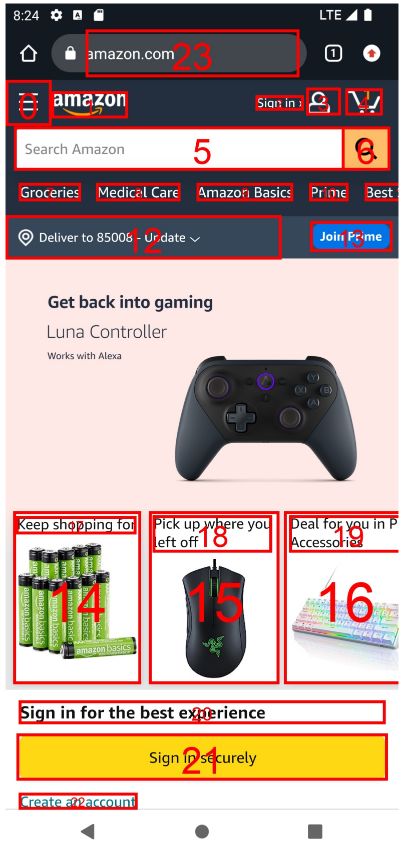
Get things done on the go.
Sign in or Sign up for free
Email, phone or Skype name
Sign later
By using this app, you agree to the terms of use
1
2
3
4

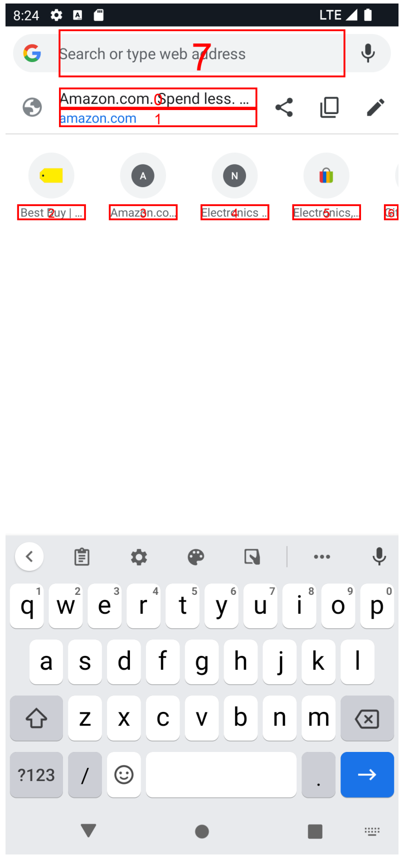
1. The whole screen -> single UI components
2. Annotate those only essential ones

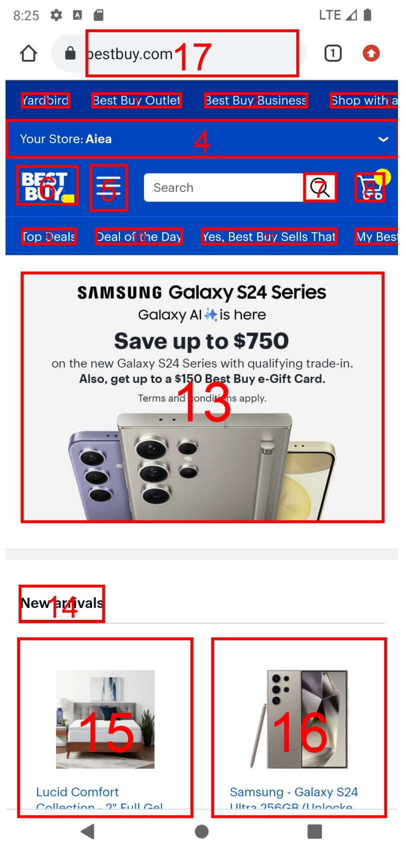
Task Completion Annotation: Examples

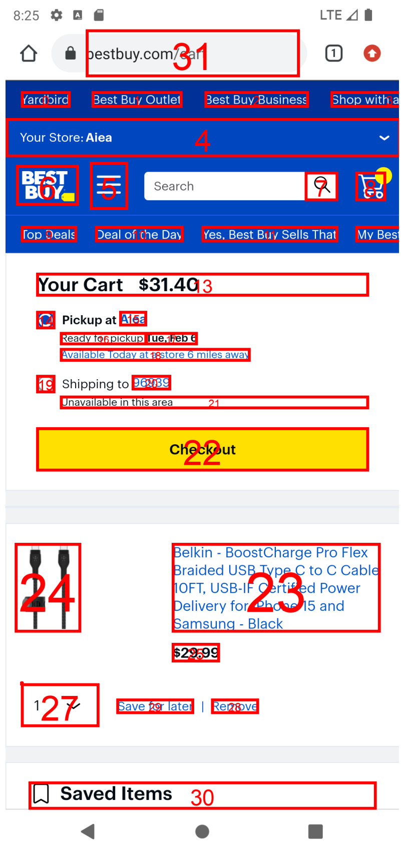
Task: Empty the shopping cart on BestBuy

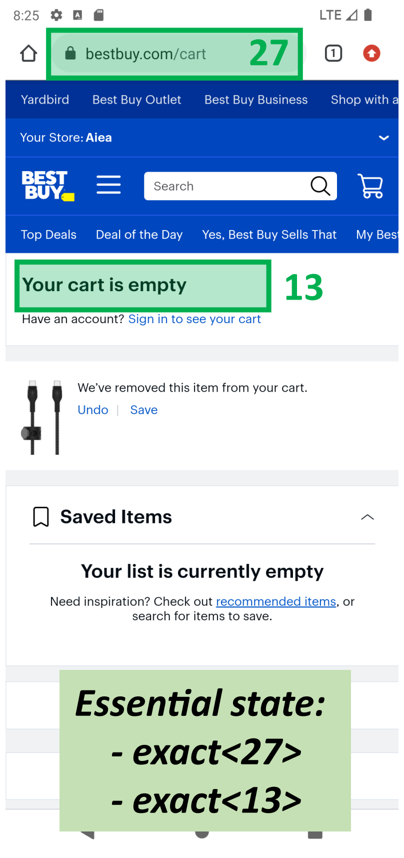










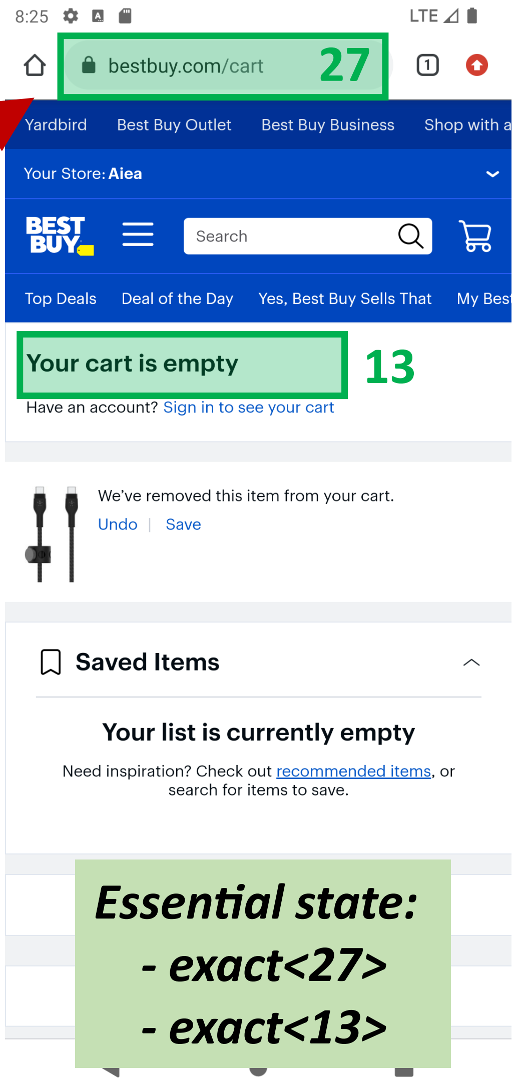


Essential state:
- exact<27>
- exact<13>

Task Completion Annotation: Examples

Task: Empty the shopping cart on BestBuy

- ❑ *Exact match* on the URL field: "bestbuy.com/cart"
- ❑ *Exact match* on the UI component with text "Your cart is empty"
- ❑ All others (actions, UI components) are omitted.


8:25 LTE

bestbuy.com/cart

Your cart is empty

Have an account? Sign in to see your cart

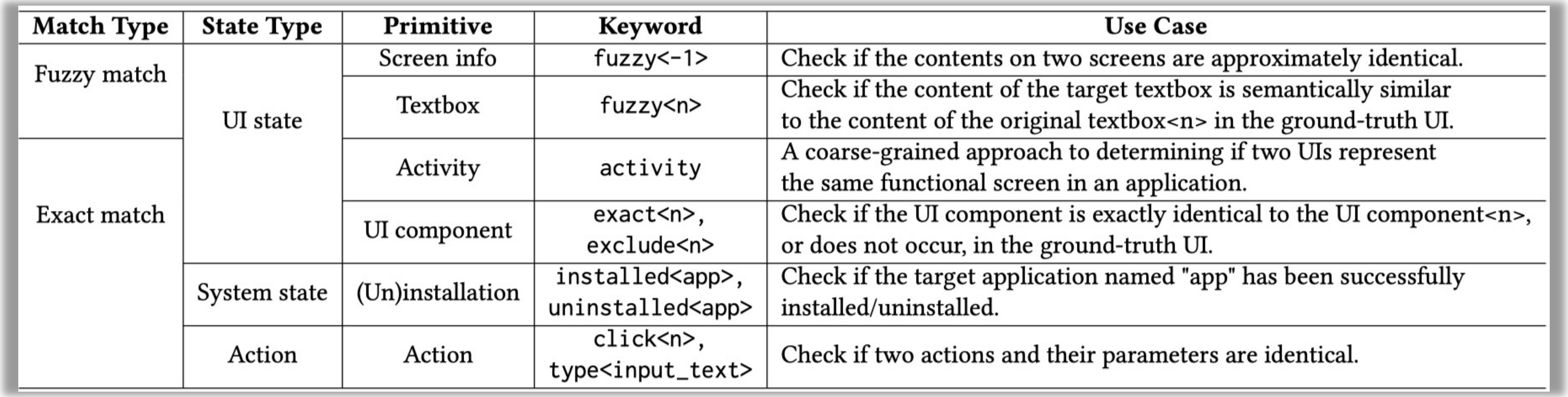
We've removed this item from your cart.

Your list is currently empty

Need inspiration? Check out recommended items, or search for items to save.

Essential state:
- exact<27>
- exact<13>

Annotation Primitives

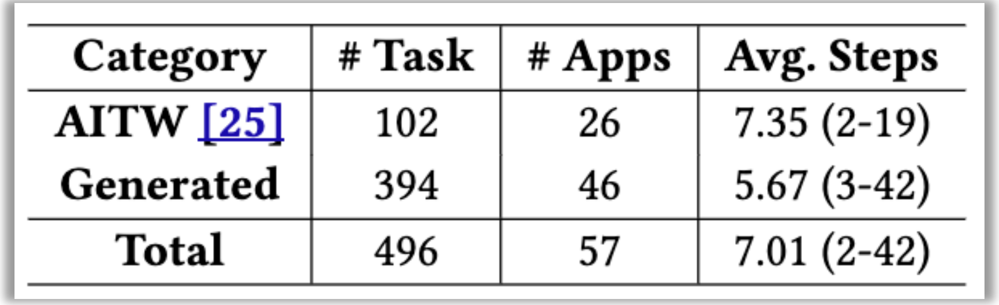


Match Type	State Type	Primitive	Keyword	Use Case
Fuzzy match	UI state	Screen info	fuzzy<-1>	Check if the contents on two screens are approximately identical.
		Textbox	fuzzy<n>	Check if the content of the target textbox is semantically similar to the content of the original textbox<n> in the ground-truth UI.
Exact match		Activity	activity	A coarse-grained approach to determining if two UIs represent the same functional screen in an application.
		UI component	exact<n>, exclude<n>	Check if the UI component is exactly identical to the UI component<n>, or does not occur, in the ground-truth UI.
	System state	(Un)installation	installed<app>, uninstalled<app>	Check if the target application named "app" has been successfully installed/uninstalled.
	Action	Action	click<n>, type<input_text>	Check if two actions and their parameters are identical.

- ❑ Two types of matching design: Exact match and fuzzy match
- ❑ Annotation at different granularity: The whole screen, individual UI components, system states, actions, etc.
- ❑ Implementation of corresponding match logic of these primitives during evaluation

LlamaTouch Dataset

- ❑ **Dataset scale: 496 tasks**
  - ❑ 102 from Android-in-the-Wild* with essential state annotated
  - ❑ 394 new-constructed ones, covering diverse daily apps; annotate from stratch



Category	# Task	# Apps	Avg. Steps
AITW [25]	102	26	7.35 (2-19)
Generated	394	46	5.67 (3-42)
Total	496	57	7.01 (2-42)

*Rawles, Christopher, et al. "Android-in-the-wild: A large-scale dataset for android device control." NeurIPS 2024.

LlamaTouch Dataset

- ❑ **Dataset scale: 496 tasks**
  - ❑ 102 from Android-in-the-Wild* with essential state annotated
  - ❑ 394 new-constructed ones, covering diverse daily apps; annotate from stratch
- ❑ **Data: Each task includes**
  - ❑ Screen representations: Pixel-level screeshots, view hierarchies, Android activities
  - ❑ Actions on each screen
  - ❑ Task instructions
  - ❑ **Annotated essential states**
  - ❑ **Task setup: A global Android emulator image (with installed apps), and env setup scripts**

*Rawles, Christopher, et al. "Android-in-the-wild: A large-scale dataset for android device control." NeurIPS 2024.

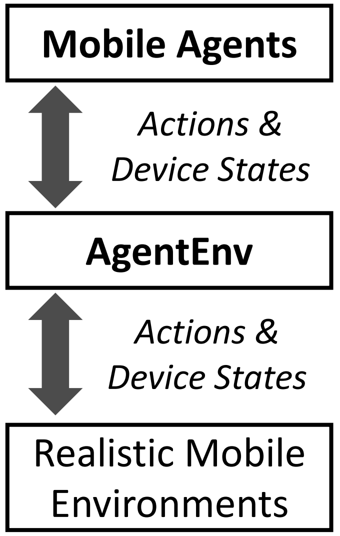
On-device Task Execution

- ❑ UI automation task in real-world environments
  - ❑ Rather than predicting actions on static datasets

Realistic Mobile Environments

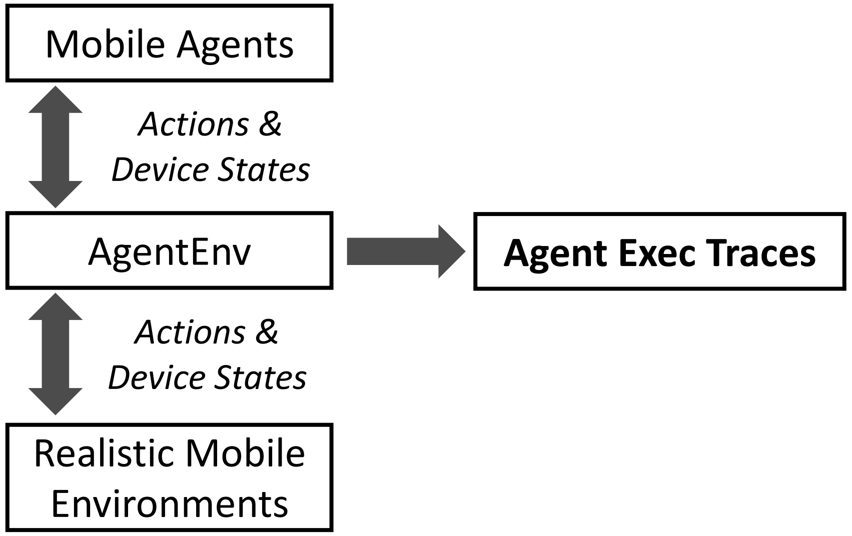
On-device Task Execution

- ❑ UI automation task in real-world environments
    - ❑ Rather than predicting actions on static datasets
- ❑ **AgentEnv:** A list of APIs to bridge mobile UI agents and real-world mobile environments



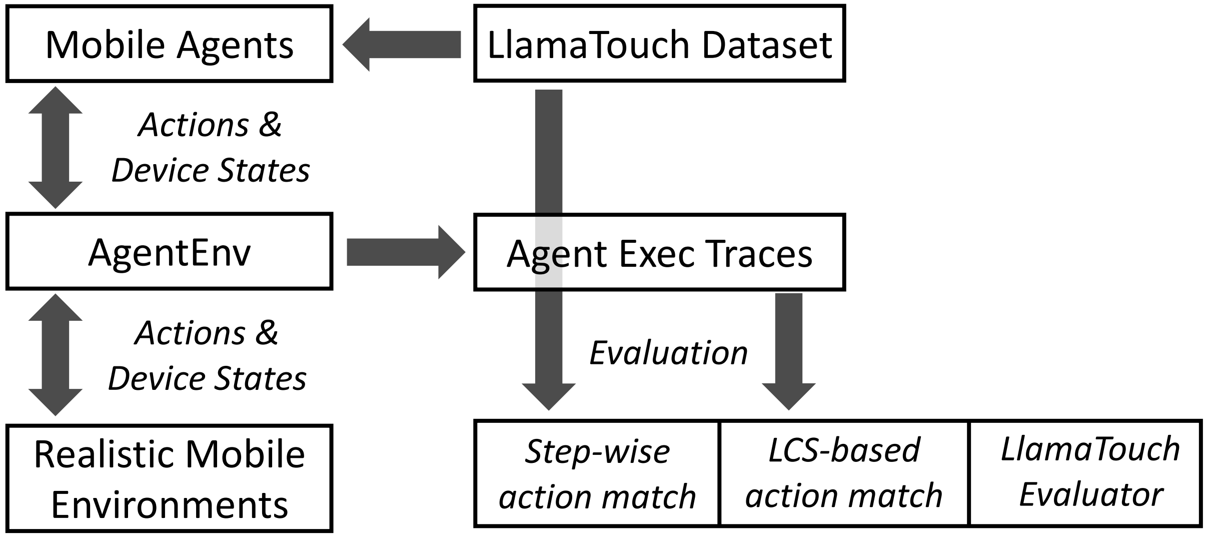
Offline Trace Evaluation

- Task execution traces are automatically logged by AgentEnv



Offline Trace Evaluation

- Task execution traces are automatically logged by AgentEnv
- **LlamaTouch Evaluator:** Compare task execution traces with predefined essential states in **LlamaTouch Dataset**



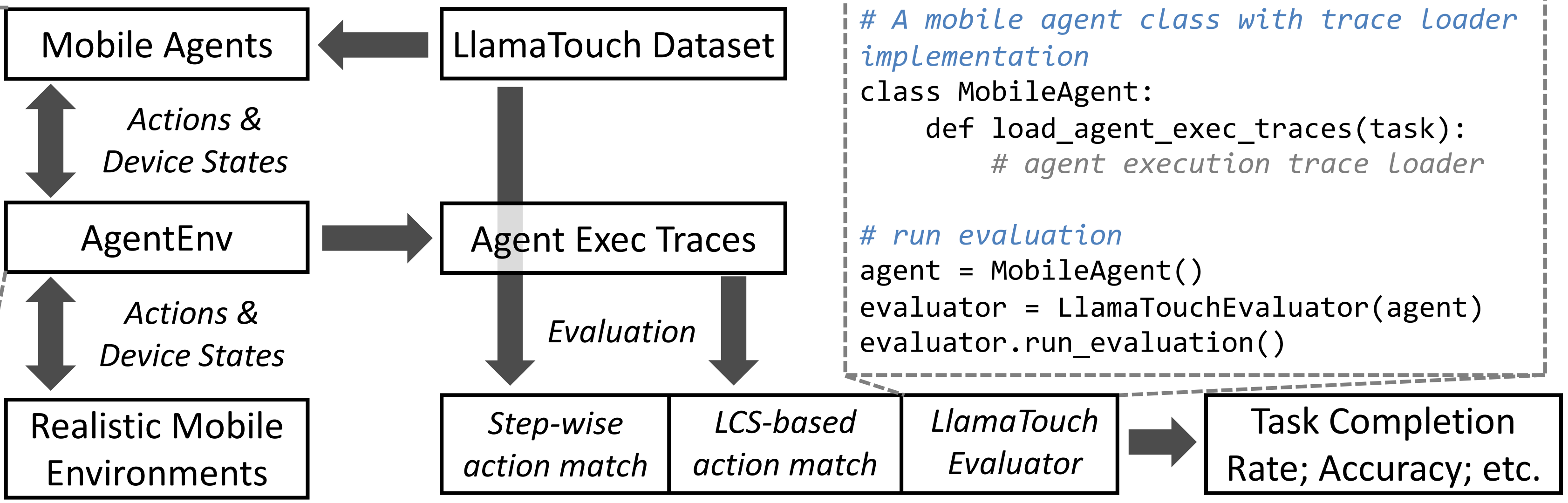
Putting Them Together

Code Demo for Mobile UI Task Execution

```
# agent/environment setup
agent = mobile_agent.init() # agent
initialization
task = agentenv.task.get()  # get task
metadata
agentenv.setup_task(task)   # task-
specific setup

# task execution
while not agent.task_complete():
    state = agentenv.get_state()
    act = agent.predict(task,
state.screenshot, state.vh)  # predict
action on current UI based on task
instruction and observed UI states
    agentenv.post(act)  # post action to
device
```

LlamaTouch Workflow



Code Demo for Trace Evaluation

```
# A mobile agent class with trace loader
implementation
class MobileAgent:
    def load_agent_exec_traces(task):
        # agent execution trace loader

# run evaluation
agent = MobileAgent()
evaluator = LlamaTouchEvaluator(agent)
evaluator.run_evaluation()
```

LlamaTouch is easy to use.

- ❑ Integrate mobile UI agents to AgentEnv
- ❑ Implement trace evaluation logic

Evaluation Setup

- ❑ **Key question: Can LlamaTouch evaluate mobile UI agents with high faithfulness?**
- ❑ **Metric: Accuracy of evaluation methods**
  - ❑ Taking human validation results as the ground truth
- ❑ **Baselines: Two action match-based evaluation methods on static datasets**
  - ❑ Step-wise action match (require two action sequences are identical)
  - ❑ Longest common subsequence (LCS)-based action match* (add non-essential actions tolerance between ground-truth actions)
- ❑ **Mobile UI agents: AutoDroid (GPT-4, MobiCom'24), Auto-UI (customized model, ACL'24), CoCo-Agent (LLaVa, ACL'24), AppAgent (GPT-4o)**

*Xing, Mingzhe, et al. "Understanding the weakness of large language model agents within a complex android environment." SIGKDD 2024.

Results from All Tasks

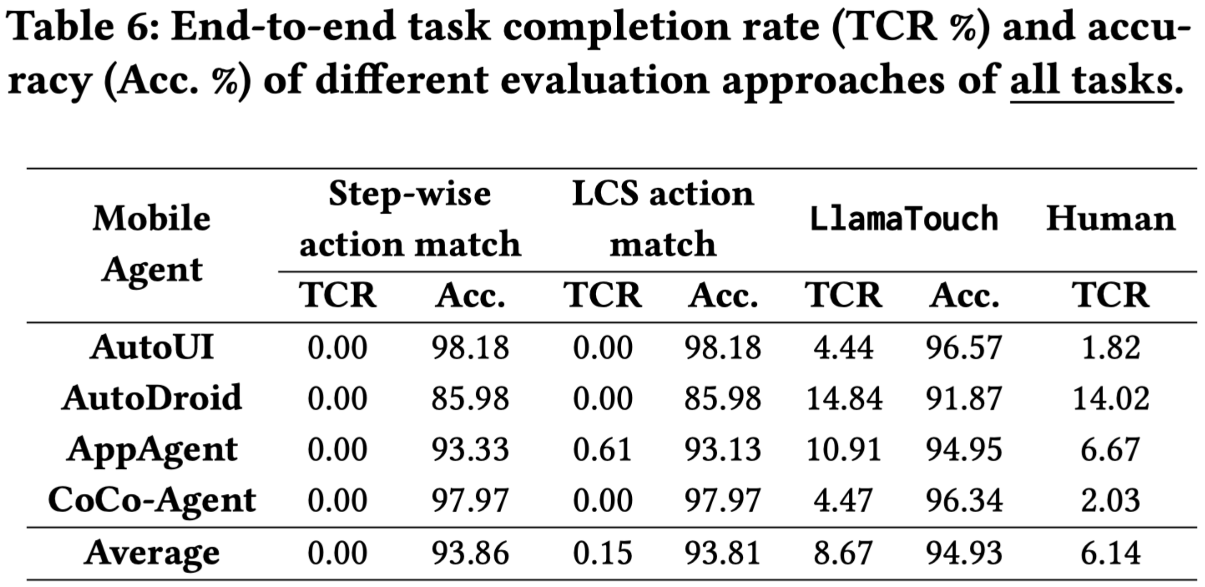
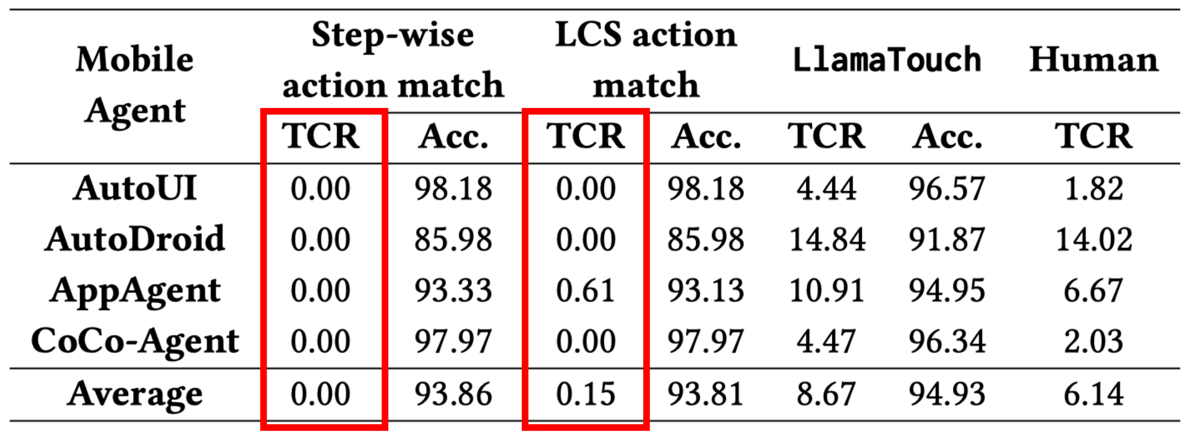


Table 6: End-to-end task completion rate (TCR %) and accuracy (Acc. %) of different evaluation approaches of all tasks.

Mobile Agent	Step-wise action match		LCS action match		LlamaTouch		Human
	TCR	Acc.	TCR	Acc.	TCR	Acc.	TCR
AutoUI	0.00	98.18	0.00	98.18	4.44	96.57	1.82
AutoDroid	0.00	85.98	0.00	85.98	14.84	91.87	14.02
AppAgent	0.00	93.33	0.61	93.13	10.91	94.95	6.67
CoCo-Agent	0.00	97.97	0.00	97.97	4.47	96.34	2.03
Average	0.00	93.86	0.15	93.81	8.67	94.93	6.14

Results from All Tasks

Table 6: End-to-end task completion rate (TCR %) and accuracy (Acc. %) of different evaluation approaches of all tasks.



Mobile Agent	Step-wise action match		LCS action match		LlamaTouch		Human
	TCR	Acc.	TCR	Acc.	TCR	Acc.	TCR
AutoUI	0.00	98.18	0.00	98.18	4.44	96.57	1.82
AutoDroid	0.00	85.98	0.00	85.98	14.84	91.87	14.02
AppAgent	0.00	93.33	0.61	93.13	10.91	94.95	6.67
CoCo-Agent	0.00	97.97	0.00	97.97	4.47	96.34	2.03
Average	0.00	93.86	0.15	93.81	8.67	94.93	6.14

- Action-match methods failed to evaluate tasks (with an almost 0% task completion rate).

Results from All Tasks

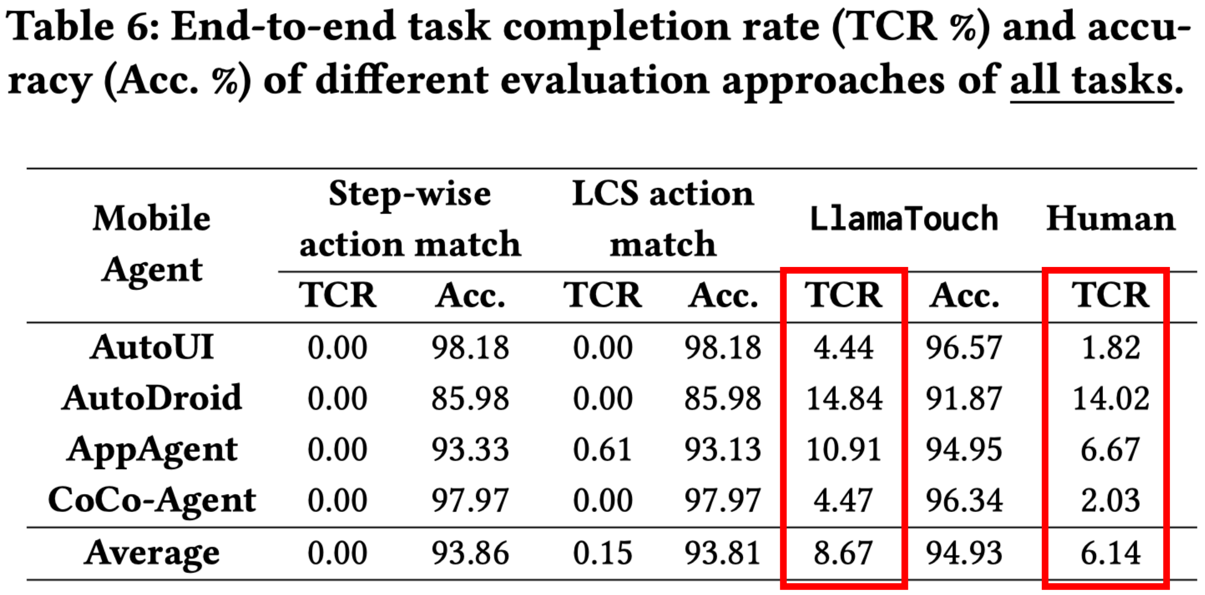


Table 6: End-to-end task completion rate (TCR %) and accuracy (Acc. %) of different evaluation approaches of all tasks.

Mobile Agent	Step-wise action match		LCS action match		LlamaTouch		Human
	TCR	Acc.	TCR	Acc.	TCR	Acc.	TCR
AutoUI	0.00	98.18	0.00	98.18	4.44	96.57	1.82
AutoDroid	0.00	85.98	0.00	85.98	14.84	91.87	14.02
AppAgent	0.00	93.33	0.61	93.13	10.91	94.95	6.67
CoCo-Agent	0.00	97.97	0.00	97.97	4.47	96.34	2.03
Average	0.00	93.86	0.15	93.81	8.67	94.93	6.14

- Action-match methods failed to evaluate tasks (with an almost 0% task completion rate).
- LlamaTouch reports task completion rates closer to human validation (e.g., 8.6 vs. 6.1).

Results from All Tasks

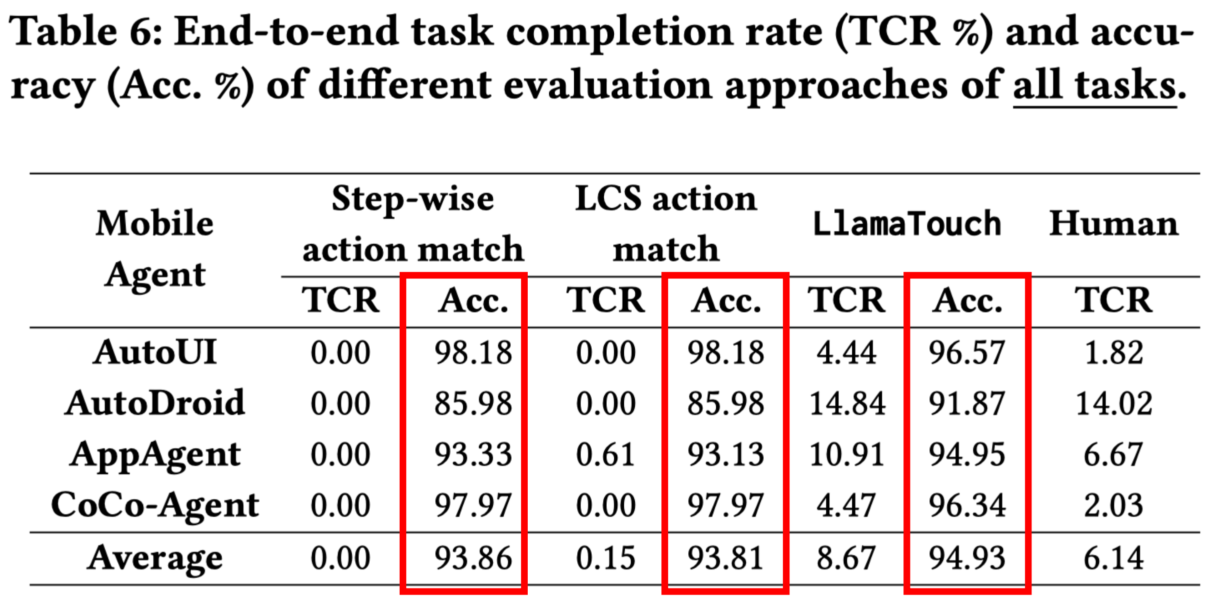


Table 6: End-to-end task completion rate (TCR %) and accuracy (Acc. %) of different evaluation approaches of all tasks.

Mobile Agent	Step-wise action match		LCS action match		LlamaTouch		Human
	TCR	Acc.	TCR	Acc.	TCR	Acc.	TCR
AutoUI	0.00	98.18	0.00	98.18	4.44	96.57	1.82
AutoDroid	0.00	85.98	0.00	85.98	14.84	91.87	14.02
AppAgent	0.00	93.33	0.61	93.13	10.91	94.95	6.67
CoCo-Agent	0.00	97.97	0.00	97.97	4.47	96.34	2.03
Average	0.00	93.86	0.15	93.81	8.67	94.93	6.14

- ❑ Action-match methods failed to evaluate tasks (with an almost 0% task completion rate).
- ❑ LlamaTouch reports task completion rates closer to human validation (e.g., 8.6 vs. 6.1).
- ❑ All methods show high accuracy as there are most false cases: UI agents cannot complete most requirements in real-world envs.

Results Completed Tasks

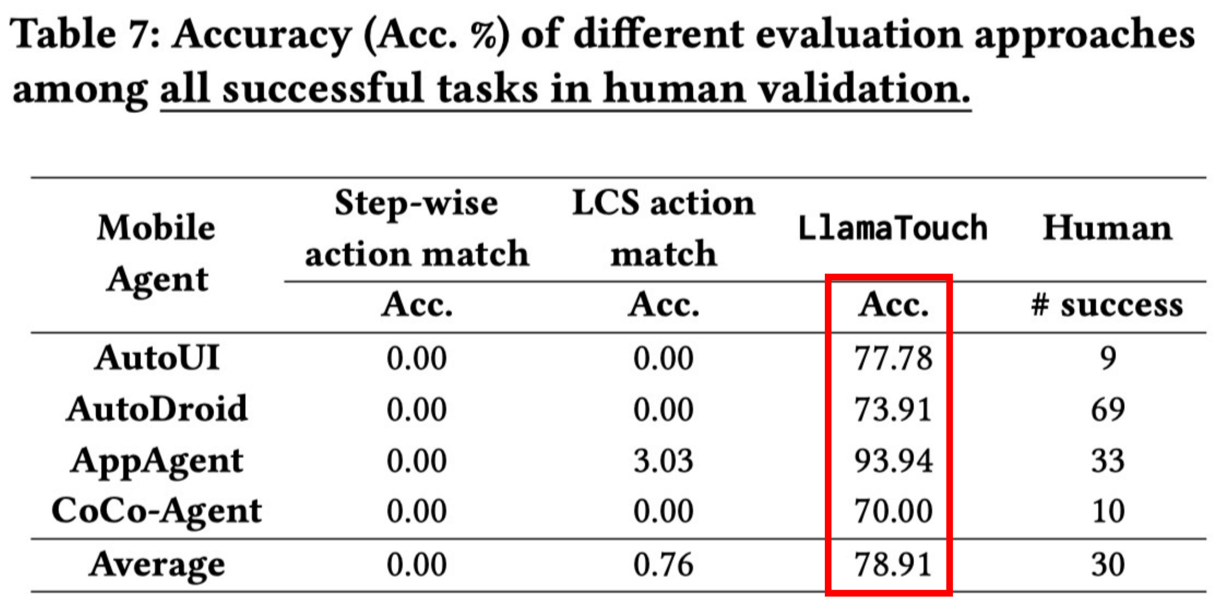


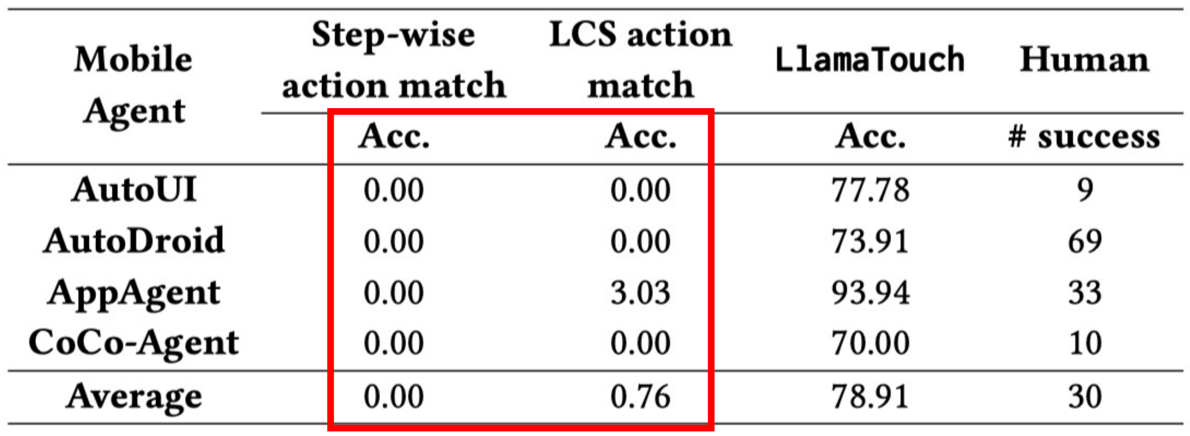
Table 7: Accuracy (Acc. %) of different evaluation approaches among all successful tasks in human validation.

Mobile Agent	Step-wise action match	LCS action match	LlamaTouch	Human
	Acc.	Acc.	Acc.	# success
AutoUI	0.00	0.00	77.78	9
AutoDroid	0.00	0.00	73.91	69
AppAgent	0.00	3.03	93.94	33
CoCo-Agent	0.00	0.00	70.00	10
Average	0.00	0.76	78.91	30

- ❑ Among all successful tasks (validated by humans), LlamaTouch achieves nearly 80% accuracy.

Results Completed Tasks

Table 7: Accuracy (Acc. %) of different evaluation approaches among all successful tasks in human validation.

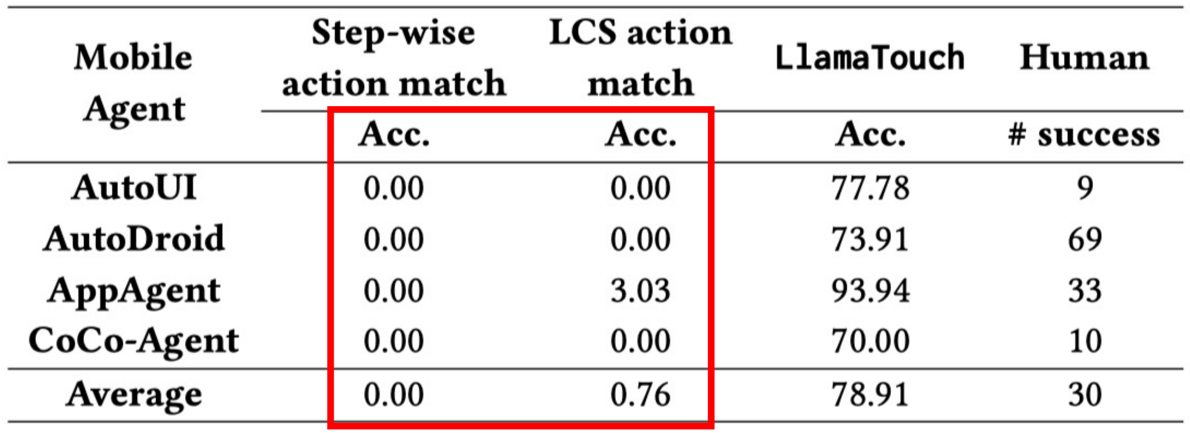


Mobile Agent	Step-wise action match	LCS action match	LlamaTouch	Human
	Acc.	Acc.	Acc.	# success
AutoUI	0.00	0.00	77.78	9
AutoDroid	0.00	0.00	73.91	69
AppAgent	0.00	3.03	93.94	33
CoCo-Agent	0.00	0.00	70.00	10
Average	0.00	0.76	78.91	30

- ❑ Among all successful tasks (validated by humans), LlamaTouch achieves nearly 80% accuracy.
- ❑ Action match approaches achieve nearly 0% accuracy.

Results Completed Tasks

Table 7: Accuracy (Acc. %) of different evaluation approaches among all successful tasks in human validation.



Mobile Agent	Step-wise action match	LCS action match	LlamaTouch	Human
	Acc.	Acc.	Acc.	# success
AutoUI	0.00	0.00	77.78	9
AutoDroid	0.00	0.00	73.91	69
AppAgent	0.00	3.03	93.94	33
CoCo-Agent	0.00	0.00	70.00	10
Average	0.00	0.76	78.91	30

- ❑ Among all successful tasks (validated by humans), LlamaTouch achieves nearly 80% accuracy.
- ❑ Action match approaches achieve nearly 0% accuracy.
- ❑ LlamaTouch significantly reduces false negative cases.

Conclusion

- 🤖 **LlamaTouch is the first faithful and scalable testbed for mobile UI task automation.**
- 😉 **Highly extensible:** New UI automation datasets, new annotation primitives, new agents, new realistic mobile environments
- 🤗 **Fully open-source:** Annotation platforms, dataset, LlamaTouch evaluator, mobile UI agents integrated into LlamaTouch

LlamaTouch is available at
https://github.com/LlamaTouch/LlamaTouch

✉️ li.zhang@bupt.edu.cn