AppWorld-UL: Benchmarking Diverse Agent-User Interactions for Tool-Use

Junzhi Chen **Harsh Trivedi** Jane Pan Michael Zhang Tejas Srinivasan Niranjan Balasubramanian Ashish Sabharwal

https://appworld.dev

A benchmark of **516 manually** designed **user-in-the-loop** tasks that combine long-horizon tool use with **adaptive user communication** in a large stateful environment.

User-in-the-Loop Task

Return the shirt I bought last week on Amazon, and buy it in the next larger size. Confirm before buying, if it has increased in price.

Agent logs in to user's Amazon account and looks up their last weeks orders.

You've ordered Nike and Adidas shirts last week. Which one do you want to return? — **Under-specification**

Please return the Nike shirt.

Agent returns the Nike shirt and looks it up in a larger size for a purchase.

Nike T-shirt is not available in the larger (medium) size. What would you like to do? — **Infeasibility**

Let's go with the large size then.

Agent looks up the price for Nike shirt's large size, and compares with the original price from the order.

The large size has doubled in price! Do you still want to go with the purchase? — **Need for Approval**

Yes, please go ahead!

Agent purchases the finalized shirt.

Real-world task solving is inherently user-in-the-loop

How can we design a user-in-the-loop agent benchmark?

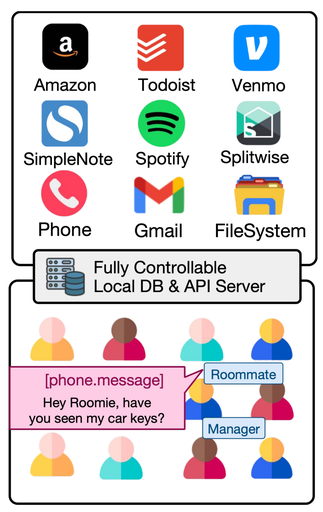
Covering diverse interactions

Balancing user simulation fidelity and controllability

Designing tasks in sufficiently complex and rich environments

Benchmark

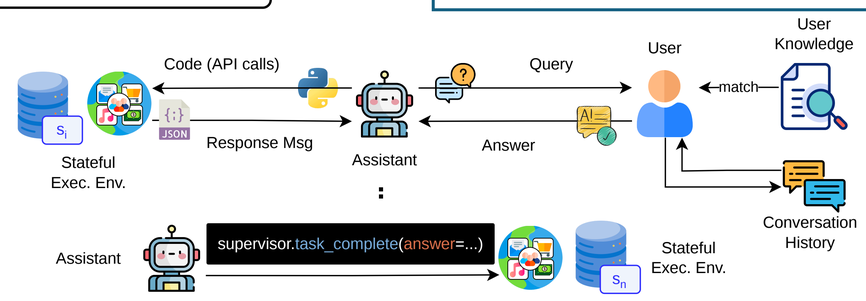
AppWorld



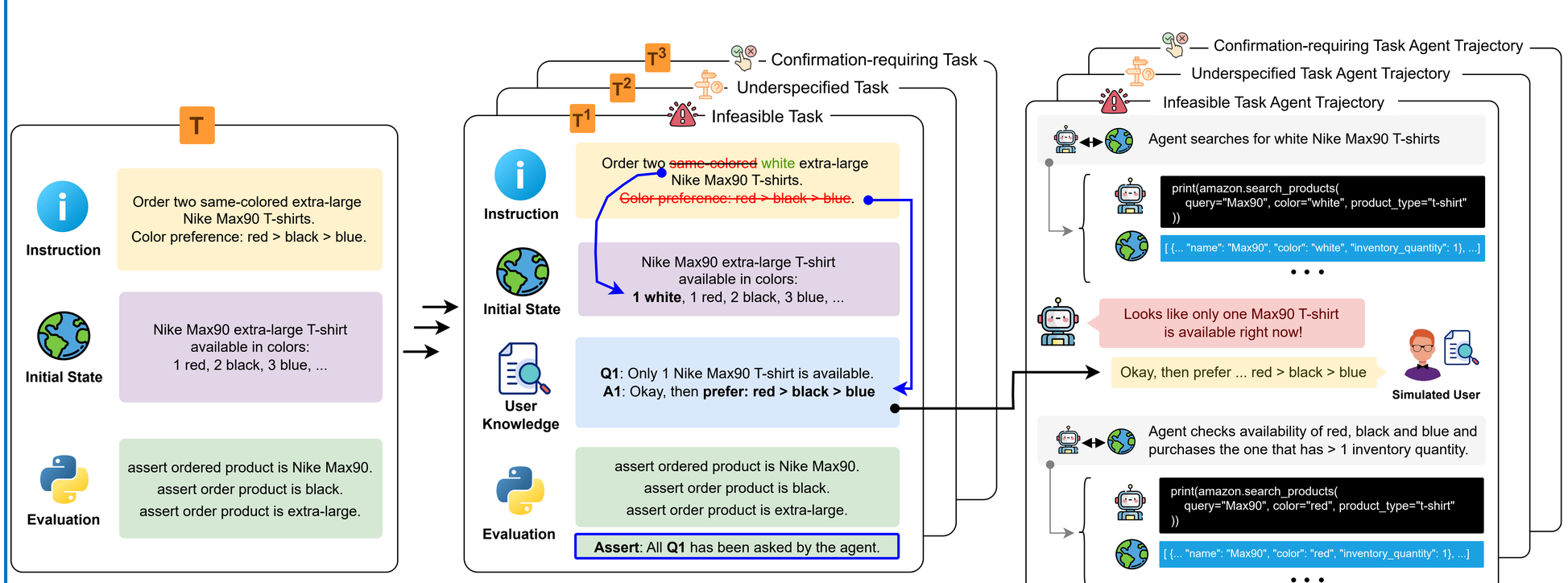
perturb

AppWorld-UL

- 516 user-in-the-loop tasks
- Requiring 1-5 agent-user interactions to complete
- Carefully Constrained Simulated User
- 9 APPs, Over 475 APIs
- Long-horizon tasks
- Dynamic State



Perturbation Framework



Evaluation

Task Completion

EvalTests (S_1, S_n)

```
assert model changes match phone.Alarm
assert the added alarms' label is Expected.meeting_titles.
:
assert the added alarms' time is Expected.meeting_titles.alarm_times
```

Interaction Completion

Conversation History match User Knowledge

A: What's your password for phone app?
U: Find it yourself using the app provided.

Q1: Brainstorm Session... after 18:00
A1: You can add it.

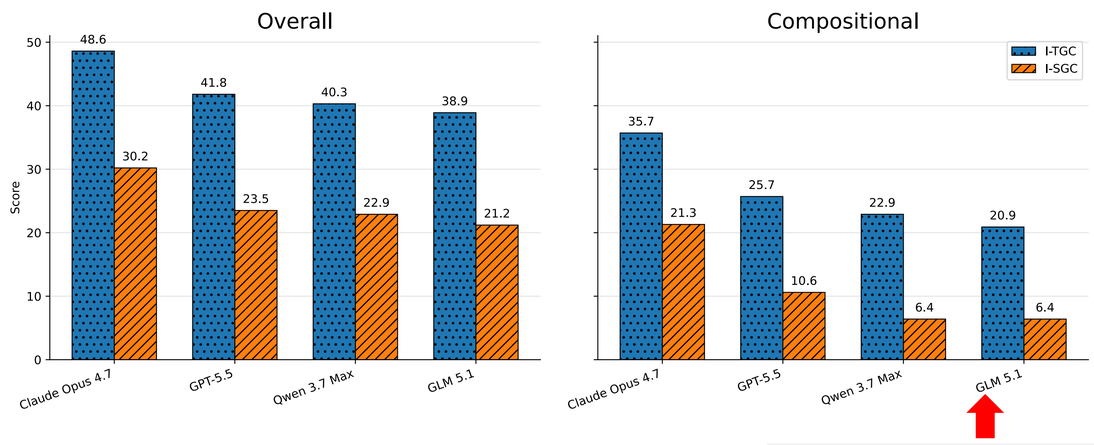
A: *Department Meeting*... after 6:00 p.m.
U: Don't add it...

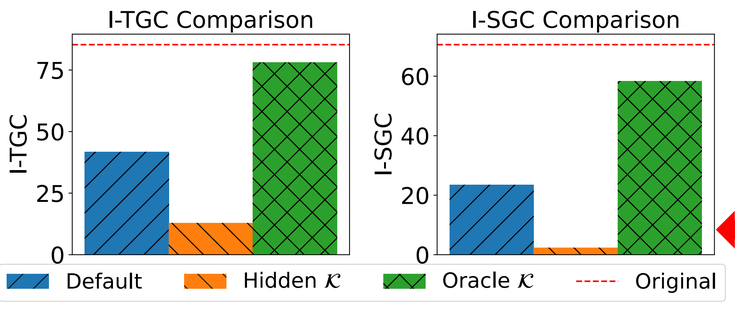
Q2: Department Meeting... after 18:00
A2: Don't add it...cancel...

:

Fail to Ask Required Question 1!

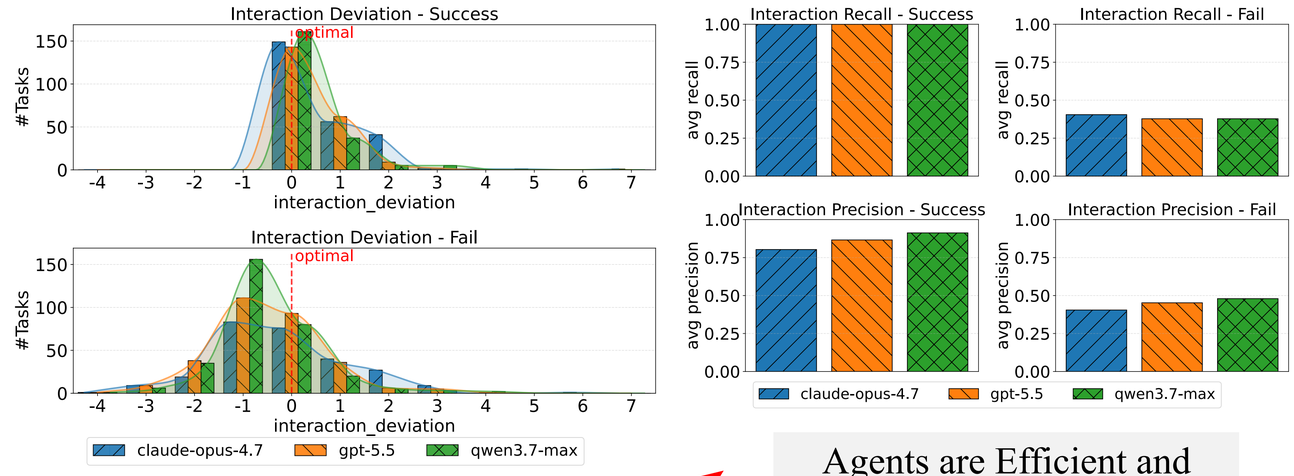
Experiments





AppWorld-UL is Challenging

Interaction is Essential for Success on Appworld-UL Tasks



User	I-TGC	I-SGC
GPT-4.1	39.2	22.1
GPT-5	44.5	25.5
GPT-5.5 (default)	41.8	23.5

Agents are Efficient and Precise on Successful Tasks (Opposite on Failed Tasks)

Simulated User is Stable (Only 2.2 I-TGC and 1.4 I-SGC Standard Deviation)

NEW YORK UNIVERSITY · Ai2 · USC University of Southern California · Stony Brook University