Designing, Evaluating, and Improving LLM Agents as Simulated Users

Yuxuan Lu

Northeastern University | https://yuxuan.lu | lu.yuxuan@northeastern.edu

Jun 2026

Introduction

About Me

Yuxuan (Leo) Lu `https://yuxuan.lu`
Ph.D. student, Northeastern University (Khoury) — advised by Prof. Dakuo Wang

Industry Experience: Applied Scientist Intern at Amazon (Sep 2024–present); ML Research Intern at LinkedIn China (Jul 2022–May 2023).
Research Focus: I work at the intersection of NLP and HCI.

Selected Works

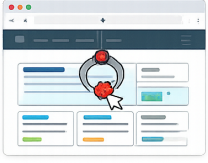
NLP / ML		**HCI**	
Finetuned LLM Agent	ACL'26	UXAgent	CHI'25 LBW
WebServ	NeurIPS'26†	+ full system	in submission
Firefly	NeurIPS'26†	Agent A/B	CHI'26 Poster

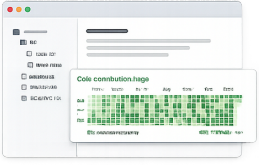
22 selected papers, 6 first-author (above; †under review).
16 co-authored — 11 accepted (ACL'26 $\times$3, ICLR'26, NAACL'24, EMNLP'23/'24, CHI'24/'25, CSCW'24/'26) and 5 under review (EMNLP'26 $\times$3, NeurIPS'26, +1 in submission).

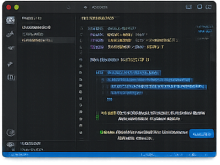
OpenClaw

opencode

Claude Code







Greeting and quick check-in

- Shorten `under review (ACL'26)` to `ACL'26 under review`
- Even shorter: `ACL'26 submission`
- If you want to keep the full phrasing, split it across two lines inside the cell
- If needed, give the right column a bit more width with weighted `X` columns

I'd recommend this version first:

```
\begin{tblr}{
  colspec={X[1.4,l]X[0.9,r]},
  row{1}={font=\bfseries}
}
  NLP & Venue/Status \\
  SFT LLM Agent & ACL'26 under review \\
  WebServ & ACL'26 under review \\
\end{tblr}
```

Why this helps:

- `ACL'26 under review` is shorter and reads more naturally on slides

Ask anything...

Build · GPT-5.4 · Default

Review 3

Session changes · Unified · Split · Collapse all

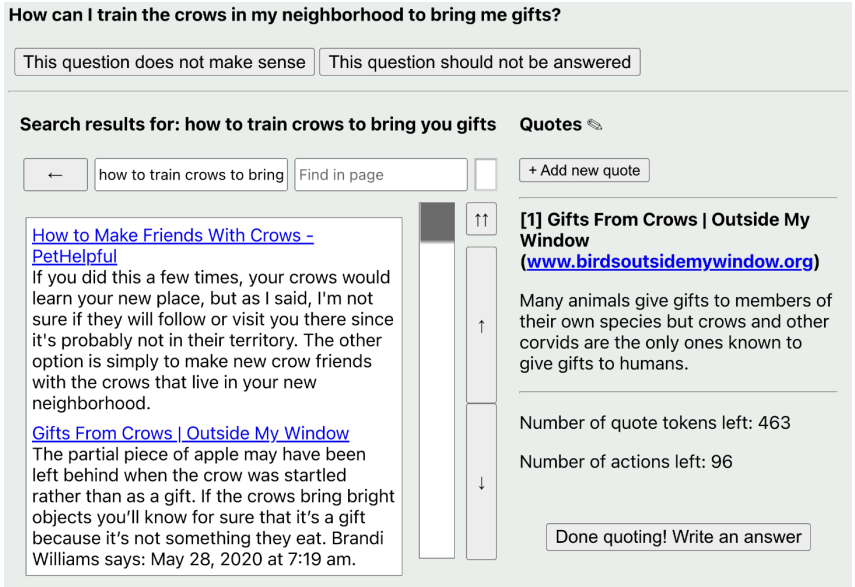
files/talks/3_papers/talk.tex +8 -1

```
{0.48\linewidth}                    {0.48\linewidth}
26 unmodified lines
204     \centering                  204     \centering
205                                 205
\includegraphics[wid                \includegraphics[widt
th=.8\linewidth]                    h=.8\linewidth]
{agents.png}                        {agents.png}
206   \end{figure}                  206   \end{figure}
207 \end{frame}                     207 \end{frame}
                                    208
                                    209 \begin{frame}[plain]
                                    210 \begin{figure}[htp]
                                    211   \centering
                                    212
                                        \includegraphics[widt
                                        h=.8\linewidth]
                                        {opencode.png}
                                    213 \end{figure}
                                    214 \end{frame}
208 }                               215 }
209                                 216
210 \begin{frame}                   217 \begin{frame}{Today's
    {Today's topic}                     topic}
211   one minute                    218   one minute seperate
```

3 Changes · All files

- files
  - talks
    - 3_papers
      - figures
        - opencode.png A
      - talk.pdf M
      - talk.tex M


16:14
5G+
Home Assistant
homeassistant-config
Ran Reload MQTT and recheck
🎉 触发成功了! 看时间戳全是 23:13:37 同一秒,reload 完 MQTT 把新阈值喂下去 → binary_sensor 翻成 ON → template trigger 评估 true → automation 跑了:
Code
binary_sensor.tesla_nav_destination_is_hom
sensor.tesla_nav_destination
sensor.tesla_nav_distance
sensor.livingroom_temperature
climate.living_room_ac
input_boolean.tesla_precool_active
你手机应该收到一条推送了 —— "Tesla Precool Cooling started..." 之类的(如果没收到,告诉我,可能 mobile_app_phone 这个 notify target 名字对不上)。
你现在车速 2 km/h(等红灯?),还剩 8.6 km。室内 25.7°C → 22°C 目标,有 ~15-20 分钟时间冷起来。
到家路径:车进 zone.home(59m 半径)→ arrive_home_scene 自动化 fire(走的是新加的 Tesla zone trigger,iPhone 不用等)→ 灯亮 + opnsense + flag → off。AC 留着,你随便管。
如果中途你改导航不回家
Add feedback...
Code
下次你想调:


How can I train the crows in my neighborhood to bring me gifts?
This question does not make sense
This question should not be answered
Search results for: how to train crows to bring you gifts
Quotes
←
how to train crows to bring
Find in page
+ Add new quote
How to Make Friends With Crows - PetHelpful
If you did this a few times, your crows would learn your new place, but as I said, I'm not sure if they will follow or visit you there since it's probably not in their territory. The other option is simply to make new crow friends with the crows that live in your new neighborhood.
Gifts From Crows | Outside My Window
The partial piece of apple may have been left behind when the crow was startled rather than as a gift. If the crows bring bright objects you'll know for sure that it's a gift because it's not something they eat. Brandi Williams says: May 28, 2020 at 7:19 am.
↑↑
↑
↓
[1] Gifts From Crows | Outside My Window (www.birdsoutsidemywindow.org)
Many animals give gifts to members of their own species but crows and other corvids are the only ones known to give gifts to humans.
Number of quote tokens left: 463
Number of actions left: 96
Done quoting! Write an answer

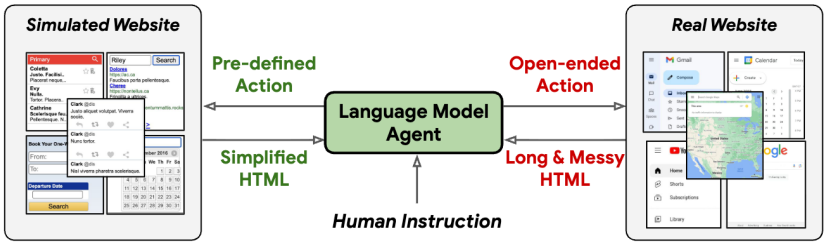
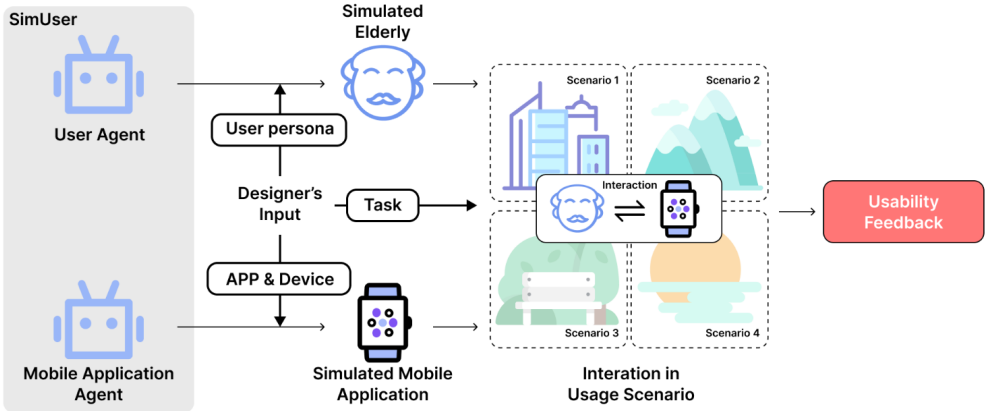


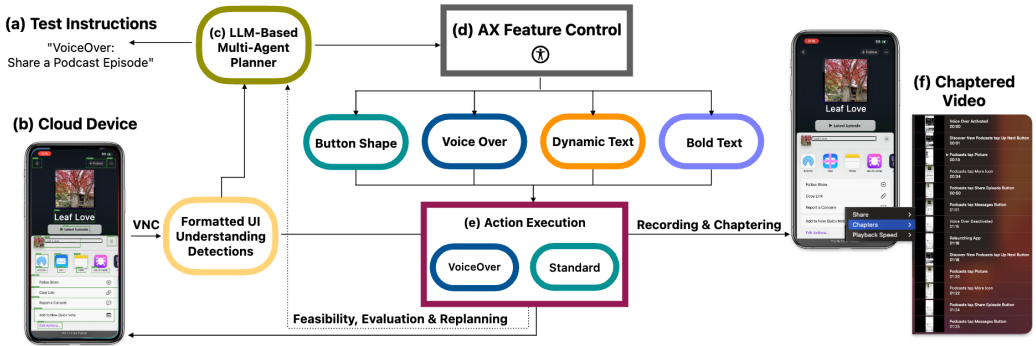
Figure 1: WebAgent, ICLR 2024

Gur et al., 2023, *The Twelfth International Conference on Learning Representations*


Joining for coffee at a cafe
KM:
AC:
[Abigail]: Hey Klaus, mind if I join you for coffee?
[Klaus]: Not at all, Abigail. How are you?
Taking a walk in the park
SM:
Arriving at school
AK:
Sharing news with colleagues
JL:
TK:
[John]: Hey, have you heard anything new about the upcoming mayoral election?
[Tom]: No, not really. Do you know who is running?
Finishing a morning routine
JM:

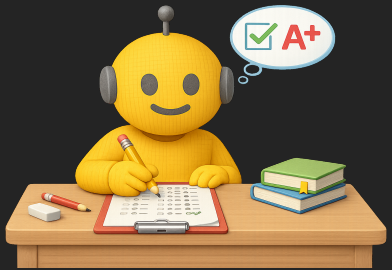


Xiang et al., 2024, *Proceedings of the CHI Conference on Human Factors in Computing Systems*

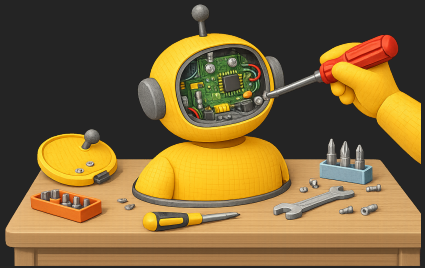

(a) Test Instructions
"VoiceOver:
Share a Podcast Episode"
(c) LLM-Based Multi-Agent Planner
(d) AX Feature Control
(b) Cloud Device
Button Shape
Voice Over
Dynamic Text
Bold Text
VNC
Formatted UI Understanding Detections
(e) Action Execution
VoiceOver
Standard
Recording & Chaptering
(f) Chaptered Video
Feasibility, Evaluation & Replanning

How can we design LLM agents as simulated users?

How can we evaluate LLM agents as simulated users?



How can we improve LLM agents as simulated users?



Case Study 1: UXAgent: A System for Simulating Usability Testing of Web Design with LLM Agents

What’s the worst nightmare for a researcher?

What's the worst nightmare for a researcher?

- Spending weeks on a study, only to find that the results are not significant
- Training a model for days, only to find a bug in your preprocessing pipeline
- The myth of "Reviewer 2"

What's the worst nightmare for a researcher?

Northeastern University

- Spending weeks on a study, only to find that the results are not significant
- Training a model for days, only to find a bug in your preprocessing pipeline
- The myth of "Reviewer 2"
- Realizing that your experiment design is flawed one week before deadline

As someone from both NLP and HCI, I think ...

Northeastern University

	NLP	HCI
Experiment Subject	Models and Machines	Human Subjects
Experiment Design	Code and Data	Study Protocol
Experiment Cost	Money	Human Participants' Time
"Debugging" Method	Code Debugging	???

Northeastern University

	NLP	HCI
Experiment Subject	Models and Machines	Human Subjects
Experiment Design	Code and Data	Study Protocol
Experiment Cost	Money	Human Participants' Time
"Debugging" Method	Code Debugging	???

Human Participants' Time is Valuable and Limited

Existing LLM Agent systems mostly works in **sandboxed environments**

Challenges

How can we better evaluate UX Research study design before running the study?

How can we better evaluate usability testing study design before running the study

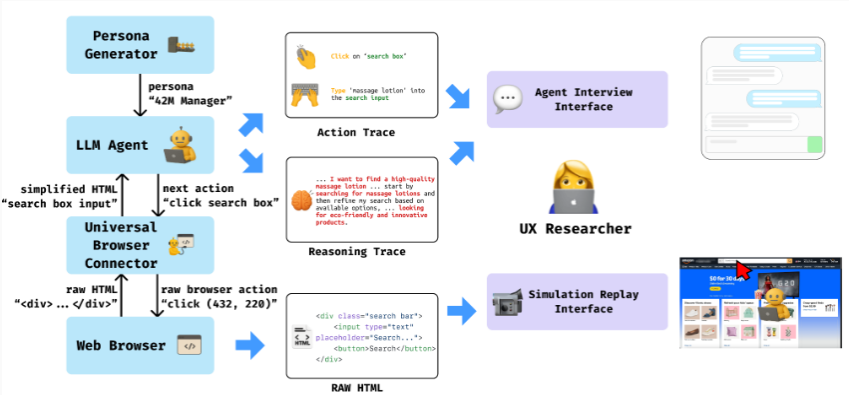
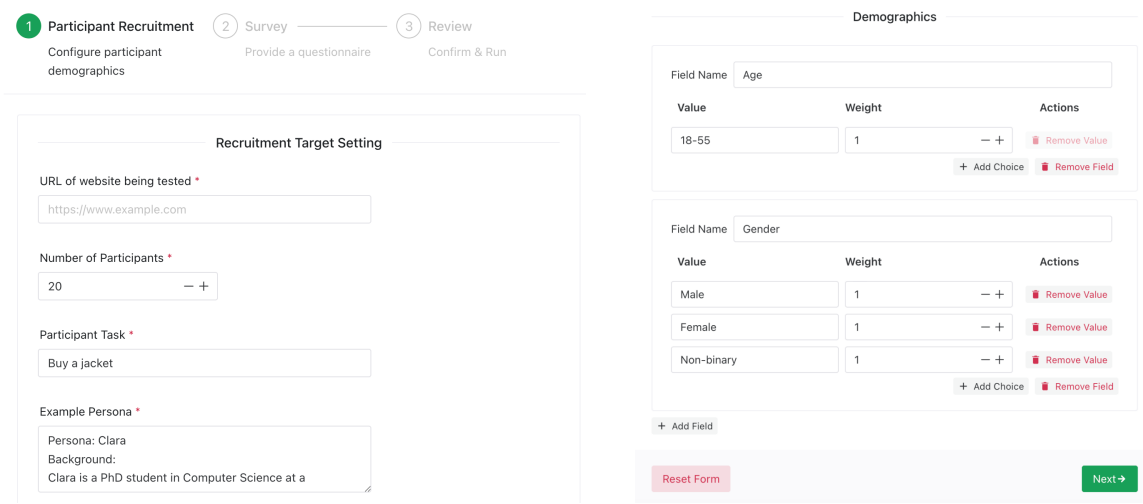


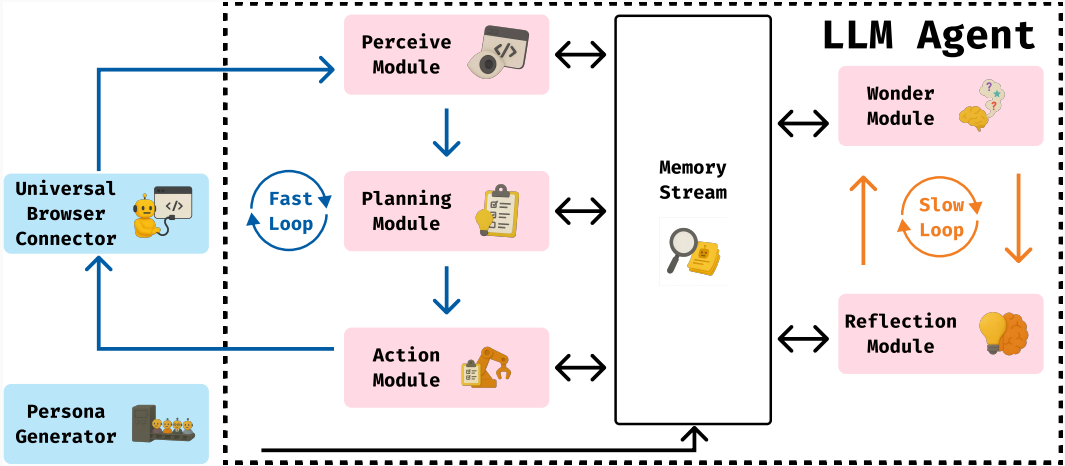
Figure 2: System Architecture of UXAgent

https://www.youtube.com/watch?v=-2xpeJ04mRA

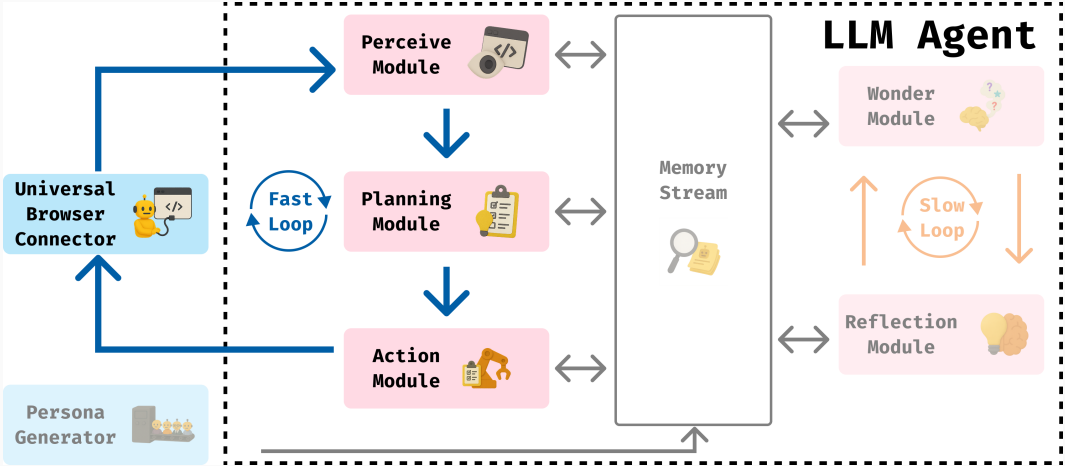


(a) Participant Task Config **(b) Demographic Distribution Config**

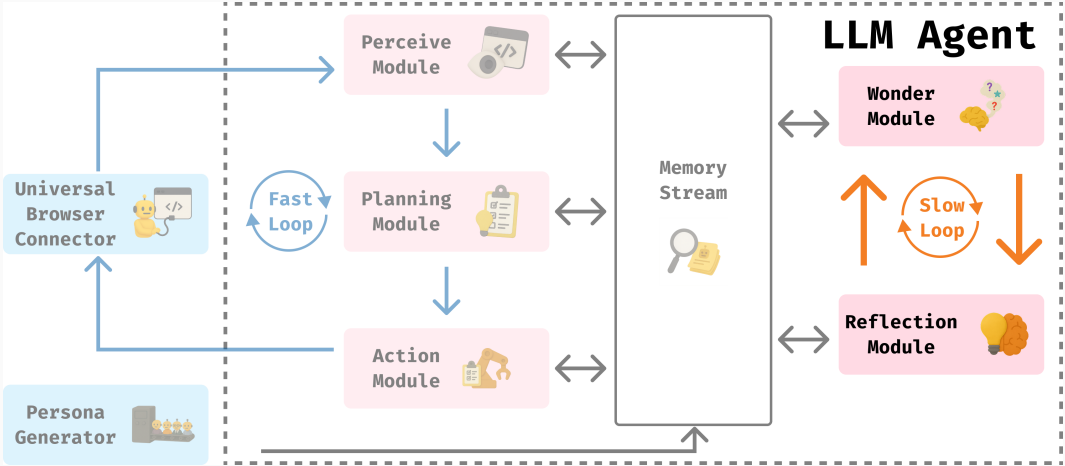
Agent Architecture



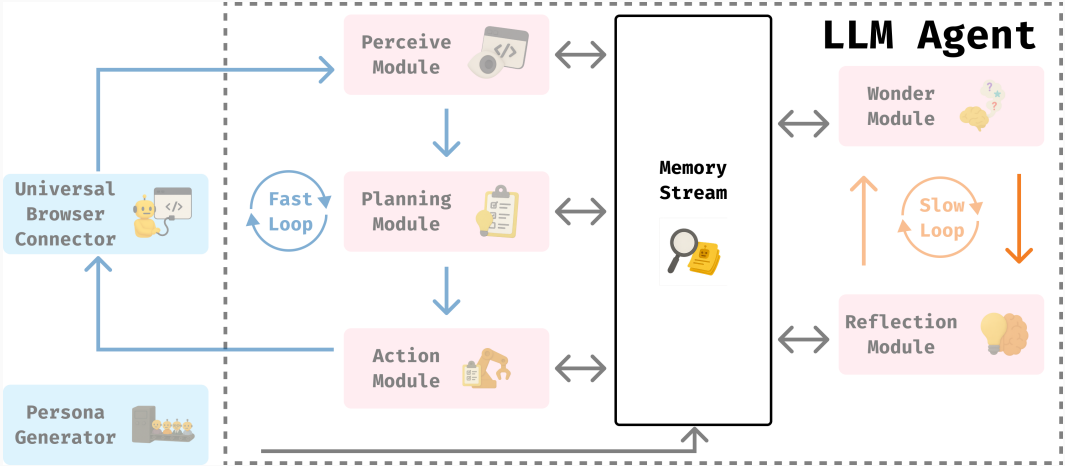
Agent Architecture - Fast Loop



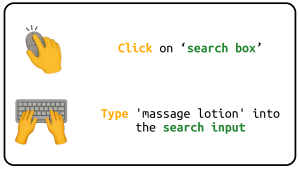
Agent Architecture - Slow Loop



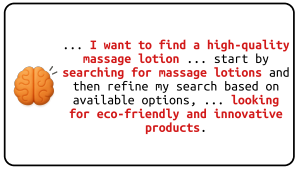
Agent Architecture - Memory Stream



Agent Architecture - Collected Data



Action Trace



Reasoning Trace

RAW HTML

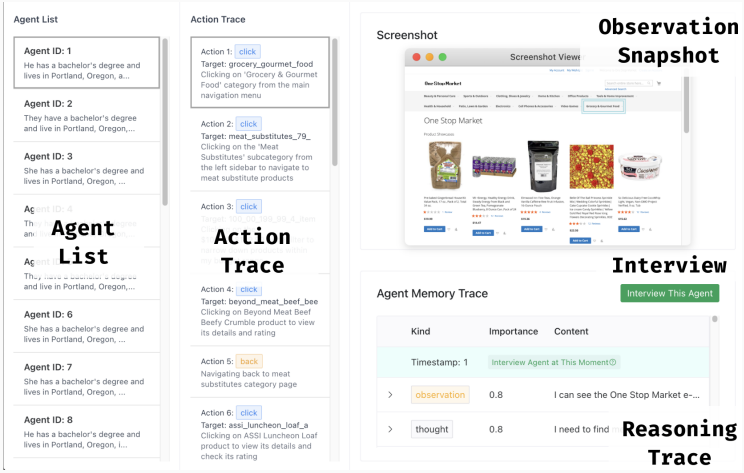


Figure 3: Result Viewer Interface

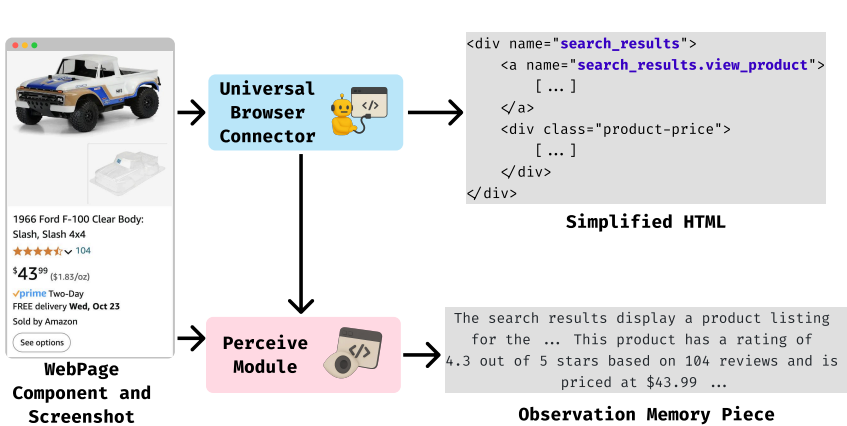


Figure 4: Universal Browser Connector

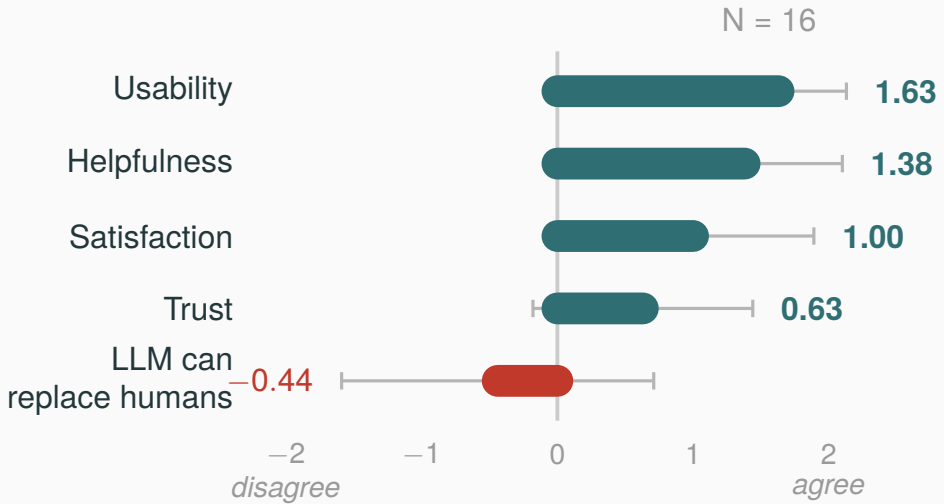
UXAgent Evaluation: can simulated agents help UX researchers?

User Study Setup

16 UX researchers · 20 simulated sessions · ~30 min / session

Task: buy a meat substitute (top-rated, $100–200), stop at checkout

Post-Study Ratings (Likert -2 to 2)



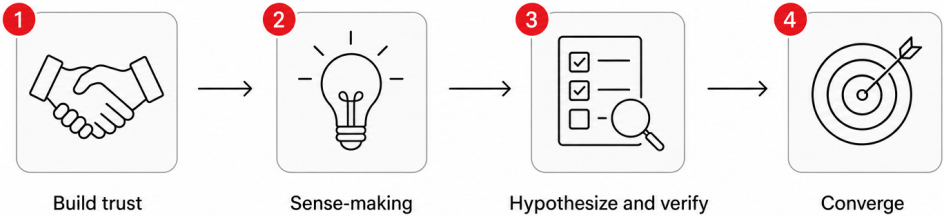
Researchers find UXAgent usable and helpful — but reject it as a *replacement* for humans.

Protocol Iteration Outcomes

14/16

proposed further study-design changes after using UXAgent

Typical Iteration Cadence



“

I trust it pretty much because it’s aligned with the reasoning.

— P3

Researchers cross-check **reasoning** ↔ **action** ↔ **screenshot** before trusting the data.
Trust rated $M = 0.62$ (SD 0.81).

Is the behavior human-like? Mixed.

convincing

"The reasoning of the actions is very reasonable."

— P7

not convincing

"I just do not think a real human participant will have actions like that."

— P6

. . . but "users always surprise me" (**P8**) — real users are unpredictable too, so some surprising agent behavior may feel *authentic* rather than less.

Northeastern University

> *I'm really afraid that researchers only rely on this and overtrust such data.*
> **— P1**

Most disagreed that agents can **replace** humans ($M = -0.44$).
Position: **augment** human studies — best for early, low-cost iteration.

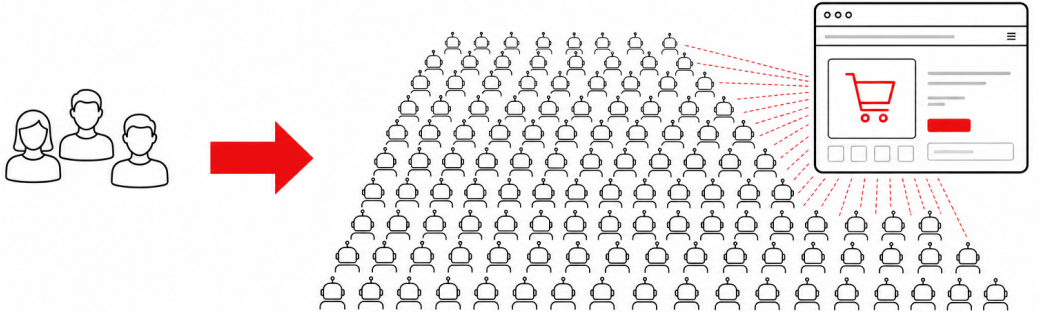
Case Study 2: Simulating Large-Scale A/B Testing Study with LLM Agents

UXAgent: LLM agents simulate users at *small scale* to help UX researchers.

UXAgent: LLM agents simulate users at *small scale* to help UX researchers.

But real design decisions ship through *large-scale A/B testing*.

Can simulated users scale up?



Three Bottlenecks in A/B Testing

No lightweight piloting

Scarce user traffic

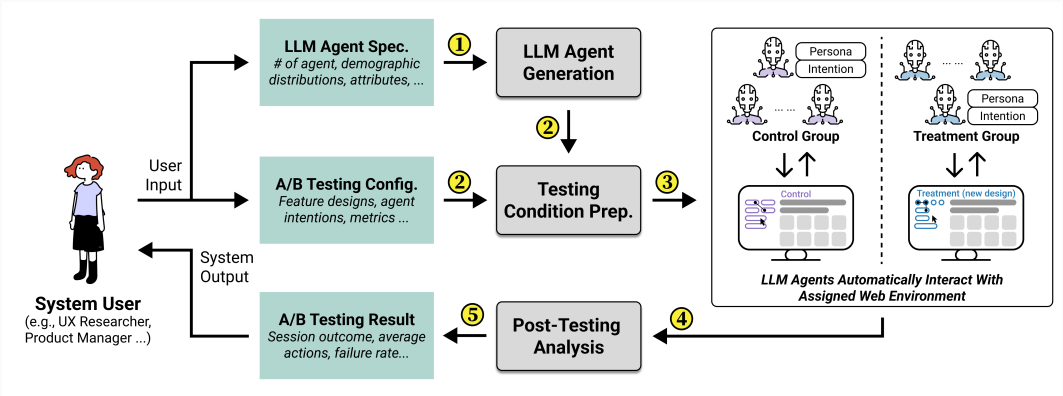
Slow feedback, weeks

From a formative study with 6 A/B-testing practitioners — many promising ideas never get piloted before launch.

Can we deploy thousands of persona-driven LLM agents on the *live* website

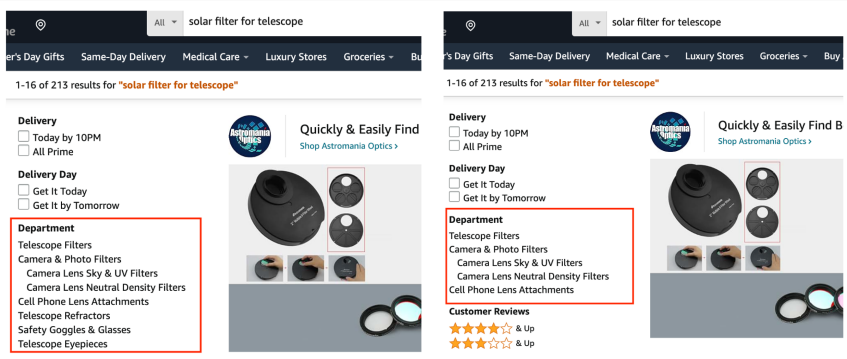
— to pilot an A/B test before spending real user traffic?

Our Solution: Agent A/B



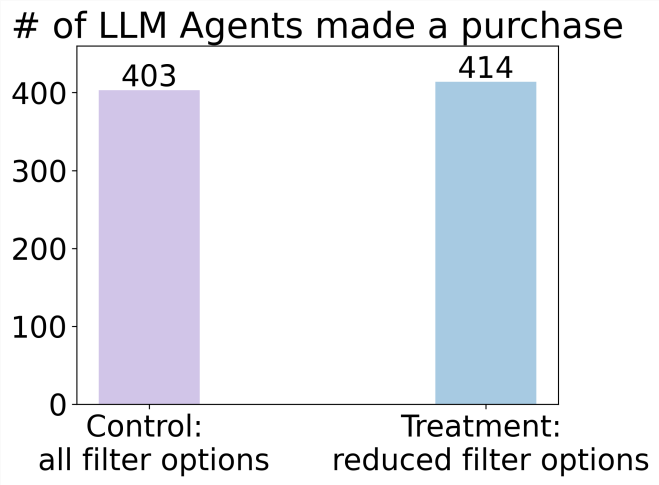
Plug-and-play with existing agent stacks (Claude computer-use, ReAct).

Case Study: Amazon.com



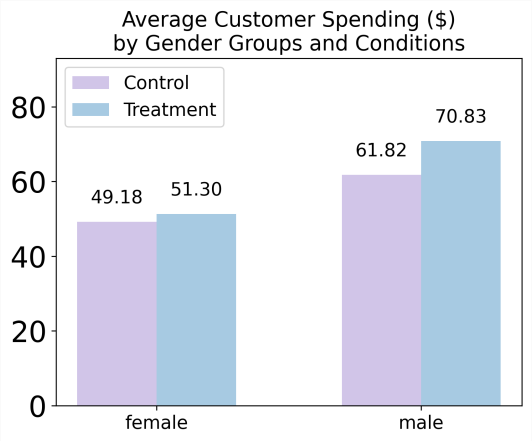
Control: full filter panel vs. **Treatment**: reduced panel (similarity ranking)
1,000 agents (500/condition) — ≈10% of a real A/B test.

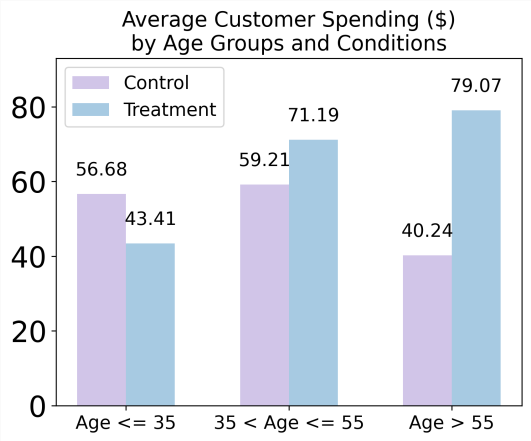
Results: Reduced Filter → More Purchases



414 vs. 403 purchases — a modest but statistically reliable increase.

Subgroup Effects





Effects are stronger for **male** and **older** personas.

Agent outcomes align *directionally* with a parallel large-scale human A/B experiment.

Simulated users can scale —
and point in the right direction.

Simulated users can scale —
and point in the right direction.

But “directionally aligned” $\neq$ “accurate”.

How faithful are these agents, really?

Case Study 3: Use Real Online Customer Behavior Data to Evaluate and Improve

Case Study 3: Use Real Online Customer Behavior Data to Evaluate and Improve

Task & Method

- We focus on the **human behavior simulation task**
  - Generate the next user action based on the context and past actions.
  - Specifically, in the online shopping scenario.

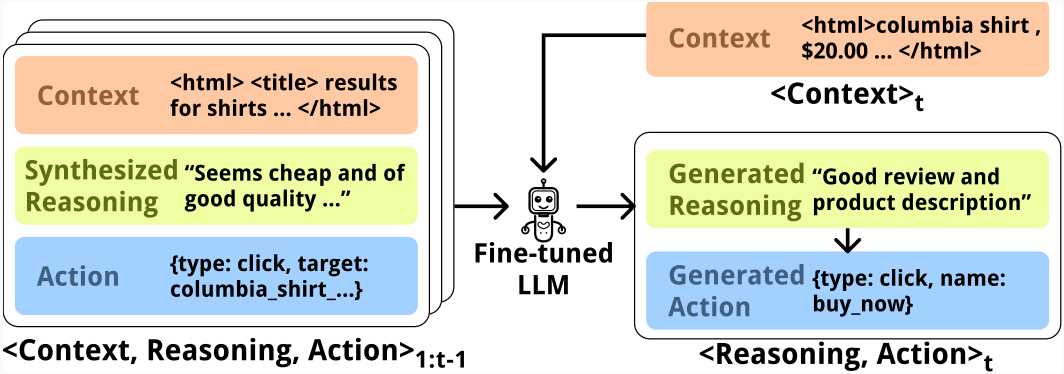
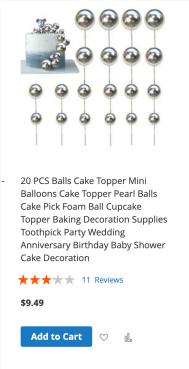


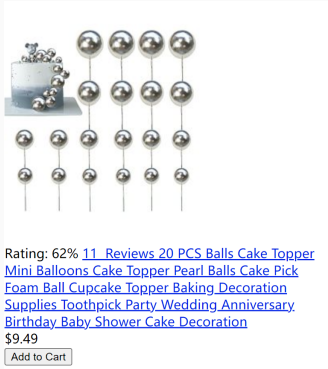
Figure 5: Overview of the next action prediction task.

Dataset

- Collected from a real-world online shopping platform.
- 31,865 sessions from 3,526 users
- 230,965 user actions
- 4,432 purchases, 27,433 terminations

Context





Rendered page

Context: simplified HTML

JS/CSS stripped, page structure kept; each interactable element gets a unique name (e.g. `product_form.add_to_cart`).

- **Action** is defined as the next raw browser action conducted by the user.
  - Generalizable to other domains beyond online shopping.
- `click` (click on an element)
- `type_and_submit` (type text and submit a form by hitting enter)
- `terminate` (user ends the session by closing the browser window)

- **Reasoning** is defined as a natural language sentence that describes the reasoning behind an action.
  - *"I want to find a comfortable piece of clothing, so I'm looking for options with high ratings."*
- Enhances the explainability of the model.
- Not present in existing datasets.
- Synthesize with teacher LLM

- To enhance LLMs' accuracy in simulating human behavior, we finetune them on the task.
  - **Input:** $\langle Context, Reasoning, Action \rangle_{1:t-1} + < Context >_t$
  - **Output:** $\langle Reasoning, Action \rangle_t$
- Training:
  - Entire session is inputed as a whole.
  - Minimize the loss of the predicted action and reasoning tokens
- Inference:
  - Input the context, past actions and corresponding reasoning.
  - Output the next action and reasoning.

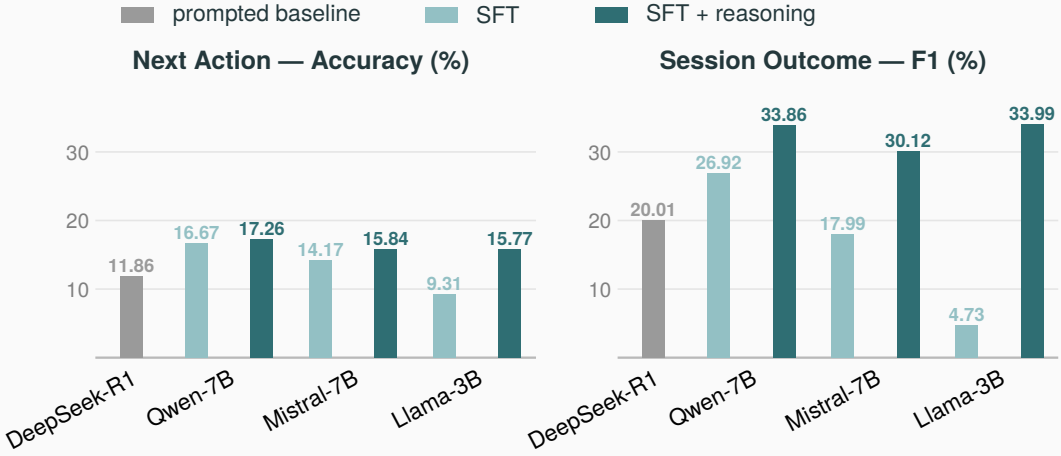
Case Study 3: Use Real Online Customer Behavior Data to Evaluate and Improve

Evaluation and Experiments

- Next Action Generation
  - Exact Match
  - Predicted action is only considered correct if both the action type (click, terminate, etc.) and the action attribute (the click target / the input text) match the ground truth.

- Next Action Generation
  - Exact Match
  - Predicted action is only considered correct if both the action type (click, terminate, etc.) and the action attribute (the click target / the input text) match the ground truth.
- Shopping Outcome Prediction
  - Essentially predicting the last action based on the session history.
  - One of `click` on a `buy_now` button or `terminate` the session.
  - F1 score

Results – Accuracy & F1



Prompting is not all you need: small fine-tuned models beat the best prompted model, and synthesized *reasoning* lifts outcome F1 further.

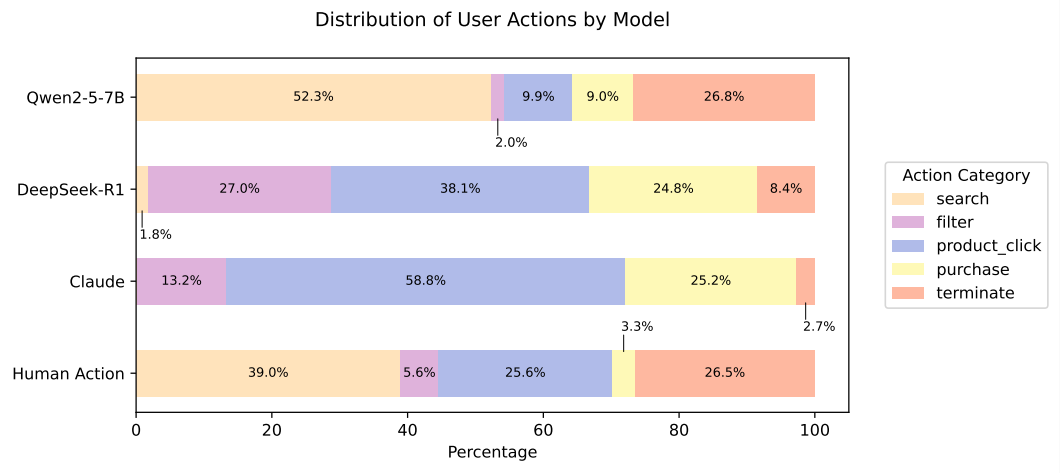


Figure 6: Distribution of the action types in the dataset.

Thank you!
Questions?