

HUMAN-AI COLLABORATION FOR UX RESEARCH

DESIGNING, EVALUATING, AND IMPROVING LLM AGENTS AS SIMULATED USERS

Yuxuan Lu

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

Mar 2026

INTRODUCTION

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.

NLP Works

SFT LLM Agent ACL'26 under review

WebServ ACL'26 under review

HCI Works

UXAgent CHI'25 LBW

Agent A/B CHI'26 Posters

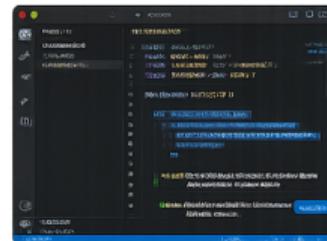
UXAgent UIST'26 under review

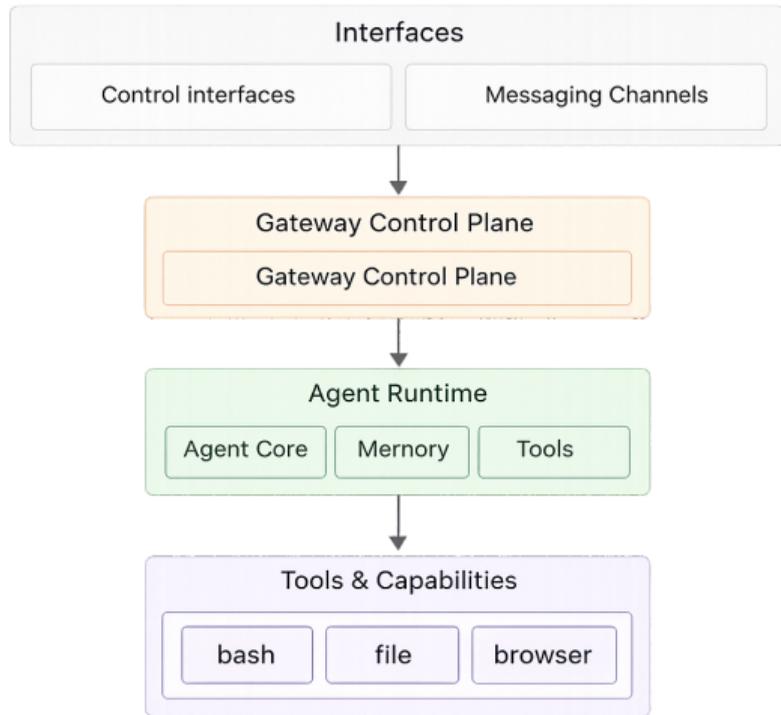
Co-authored: 3 accepted NLP papers (ICLR/EMNLP/NAACL), 4 accepted HCI papers (CHI/CSCW/UbiComp), plus 6 NLP and 1 HCI papers under review.



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[1.4,1]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
\includegraphics[wid \includegraphics[width
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[width
h=.8\linewidth]
{opencode.png}
213 \end{figure}
214 \end{frame}
208 } 208 }
209
210 \begin{frame} 210 \begin{frame}{Today's
{Today's topic} topic}
211 one minute 211 one minute separate
separate slide slide
```

3 Changes

All files

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

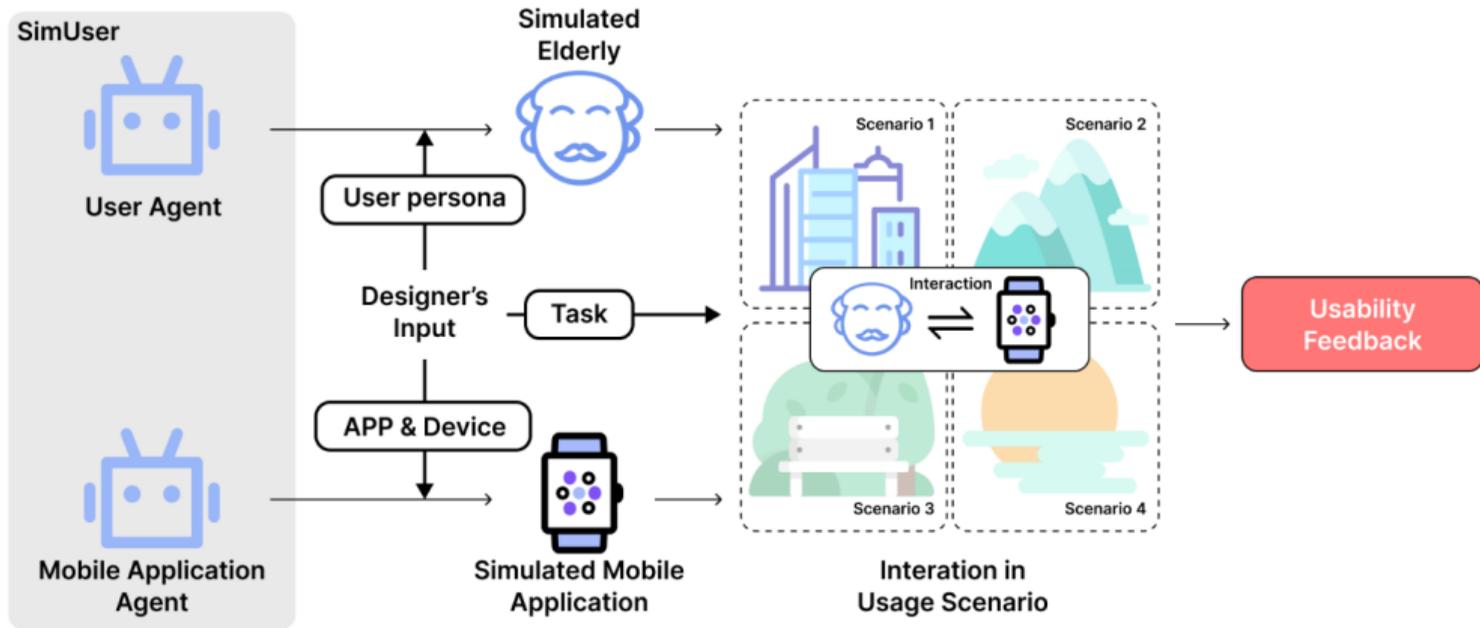
- Spending weeks on a study, only to find that the results are not significant
- Training a model for days, only to find a bug in your preprocessing pipeline
- Realizing that your experiment design is flawed one week before deadline

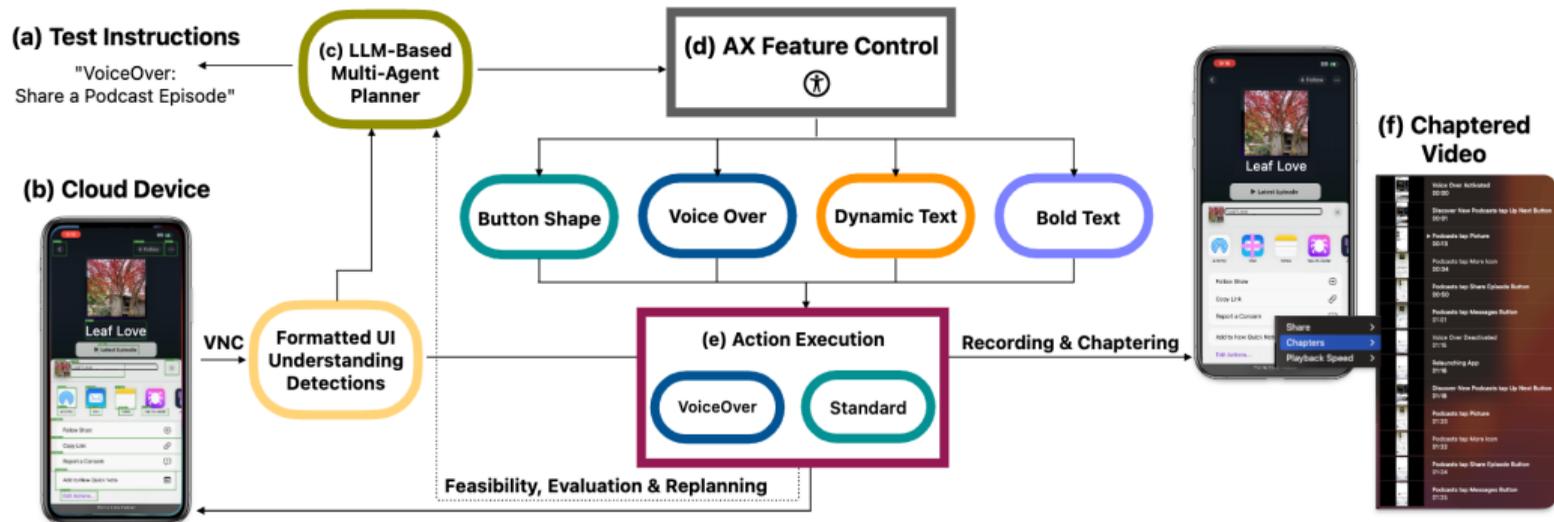
	NLP	HCI
Experiment Subject	Models and Machines	Human Subjects
Experiment Design	Code and Data	Study Protocol
Experiment Cost	Money	Human Participants' Time
"Debugging" Method	Code Debugging	???

	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







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

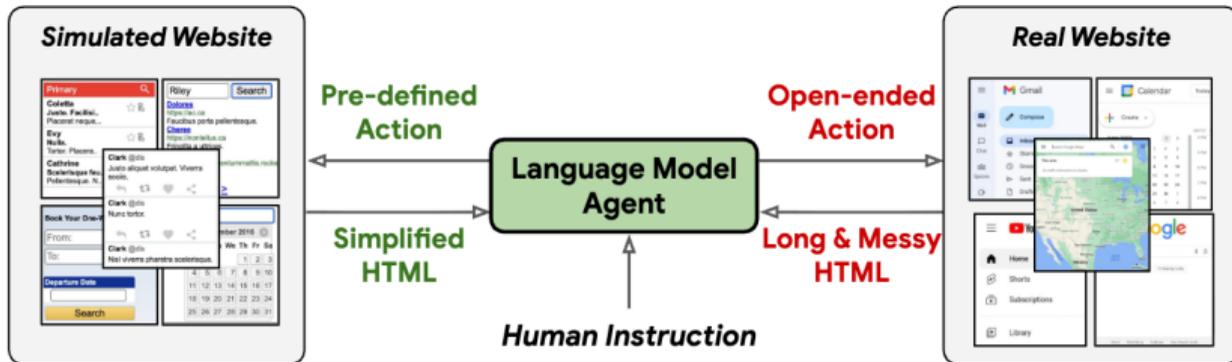


Figure 1: WebAgent, ICLR 2024



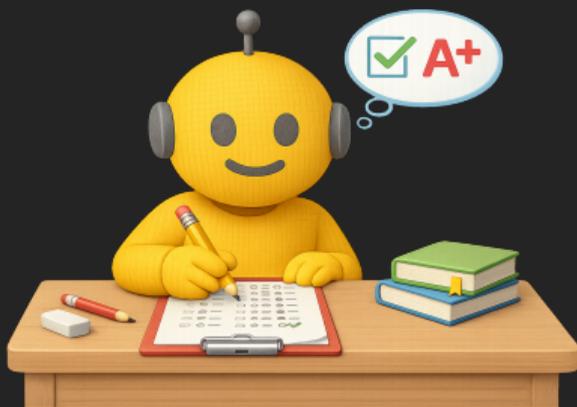
Existing LLM Agent systems mostly works in **sandboxed environments**



How can we design LLM agent systems for UX research?



How can we evaluate LLM agent systems?



How can we improve LLM agent systems?



CASE STUDY 1: UXAGENT: A SYSTEM FOR SIMULATING USABILITY TESTING OF WEB DESIGN WITH LLM AGENTS

**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**

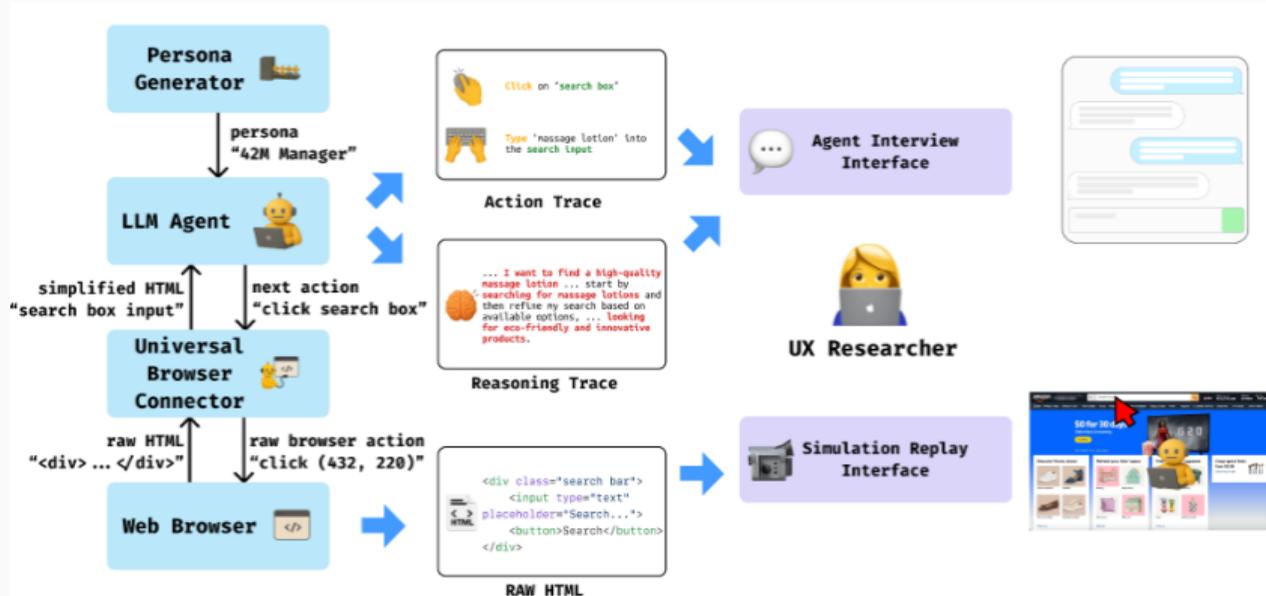


Figure 2: System Architecture of UXAgent

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

- 1 Participant Recruitment
- 2 Survey
- 3 Review

Configure participant demographics Provide a questionnaire Confirm & Run

Recruitment Target Setting

URL of website being tested *

Number of Participants *

- +

Participant Task *

Example Persona *

Persona: Clara
Background:
Clara is a PhD student in Computer Science at a

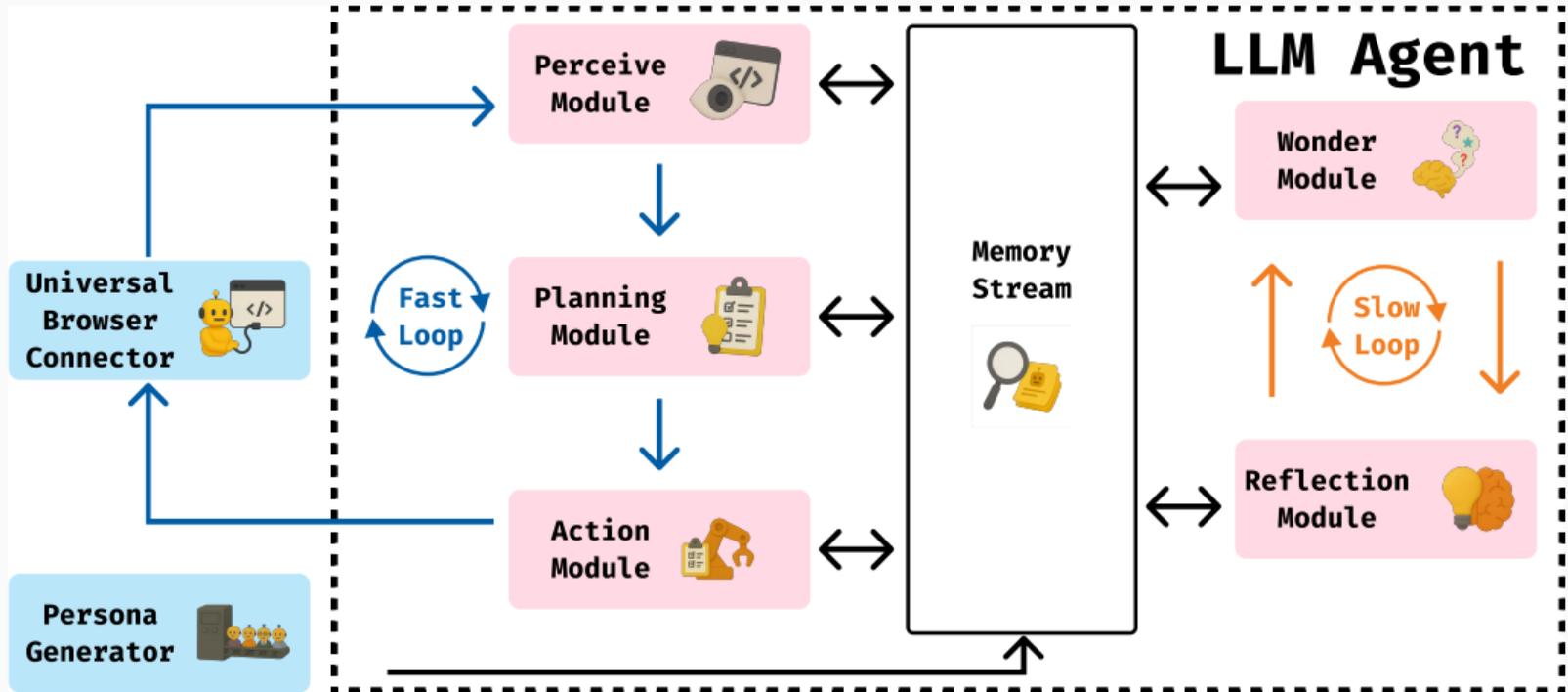
(a) Participant Task Config

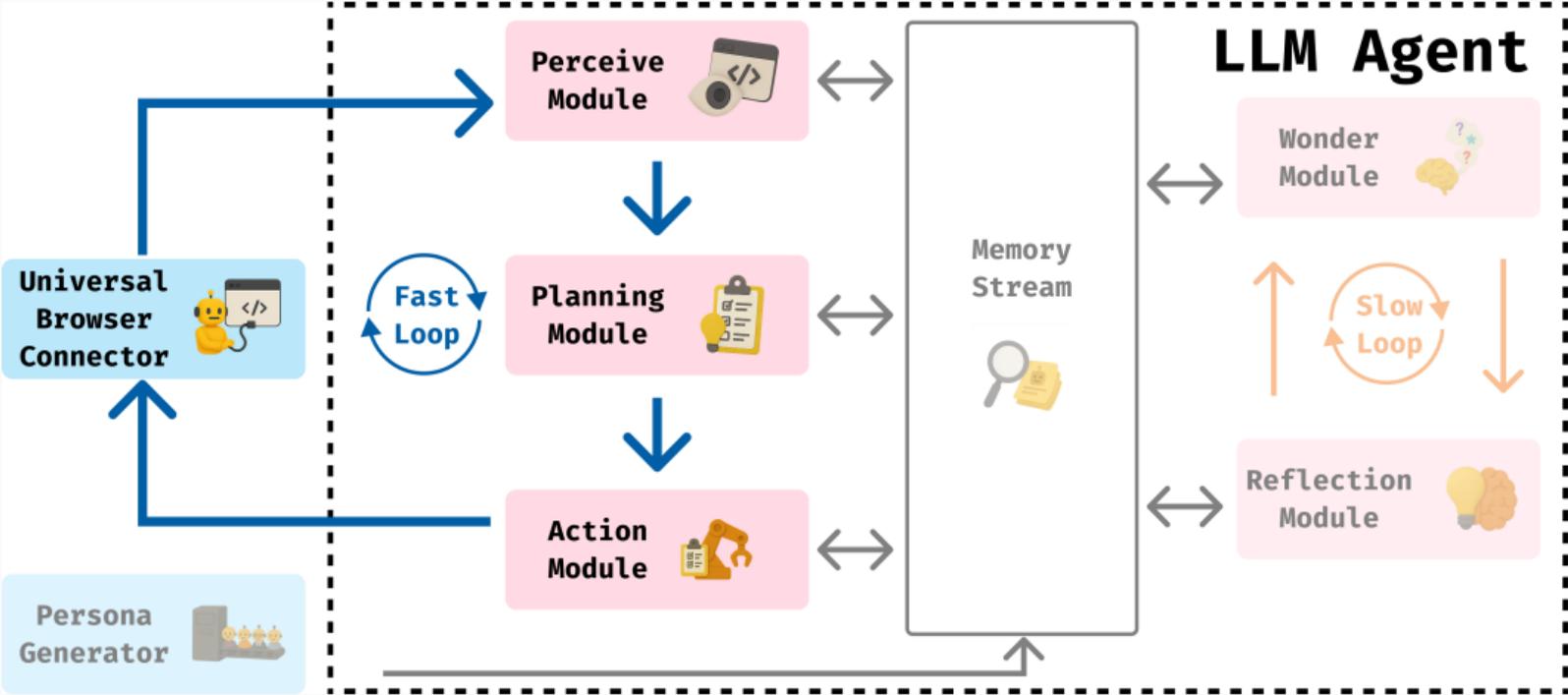
Demographics

Field Name	Value	Weight	Actions
Age	18-55	1 - +	<input type="button" value="Remove Value"/> <input type="button" value="Remove Field"/>
<input type="button" value="+ Add Choice"/> <input type="button" value="Remove Field"/>			

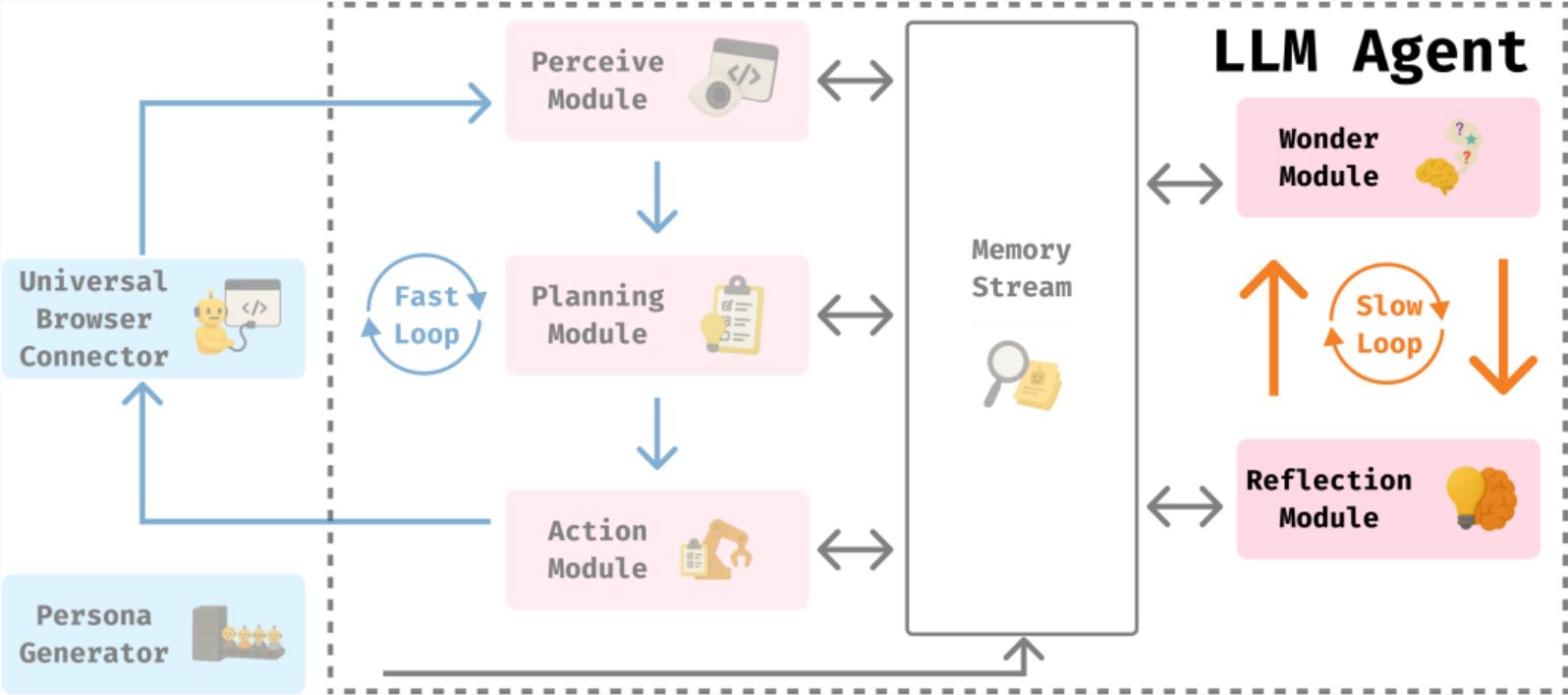
Field Name	Value	Weight	Actions
Gender	Male	1 - +	<input type="button" value="Remove Value"/>
	Female	1 - +	<input type="button" value="Remove Value"/>
	Non-binary	1 - +	<input type="button" value="Remove Value"/>
<input type="button" value="+ Add Choice"/> <input type="button" value="Remove Field"/>			

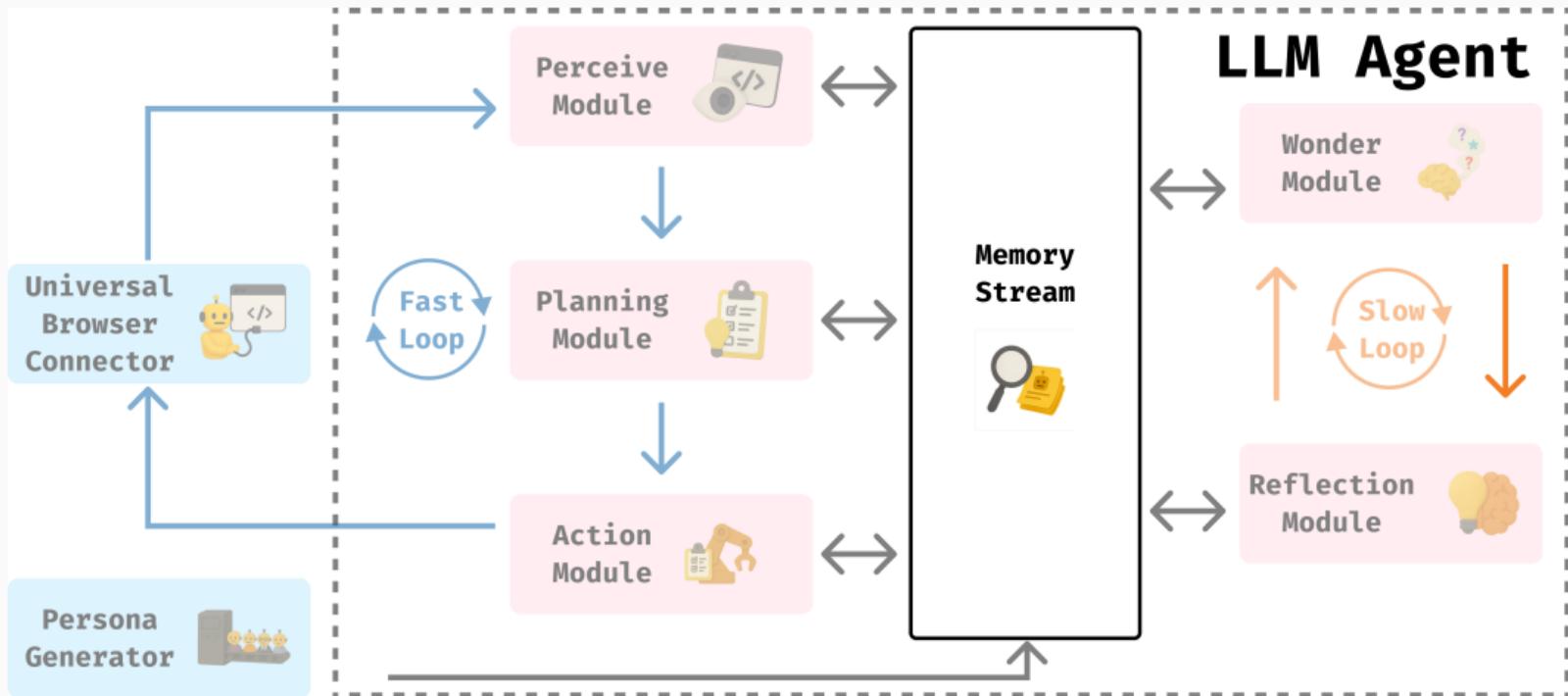
(b) Demographic Distribution Config





AGENT ARCHITECTURE - SLOW LOOP







Click on 'search box'

Type 'massage lotion' into the search input

Action Trace



... I want to find a high-quality massage lotion ... start by searching for massage lotions and then refine my search based on available options, ... looking for eco-friendly and innovative products.

Reasoning Trace



```
<div class="search bar">  
  <input type="text"  
  placeholder="Search...">  
  <button>Search</button>  
</div>
```

RAW HTML

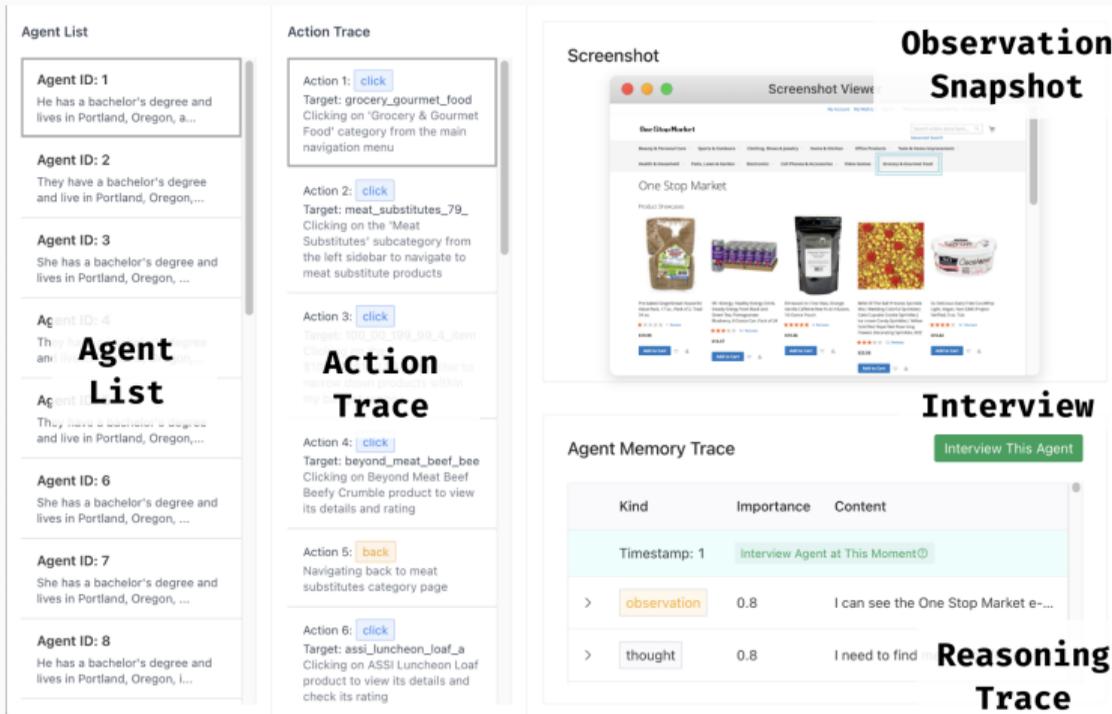


Figure 3: Result Viewer Interface

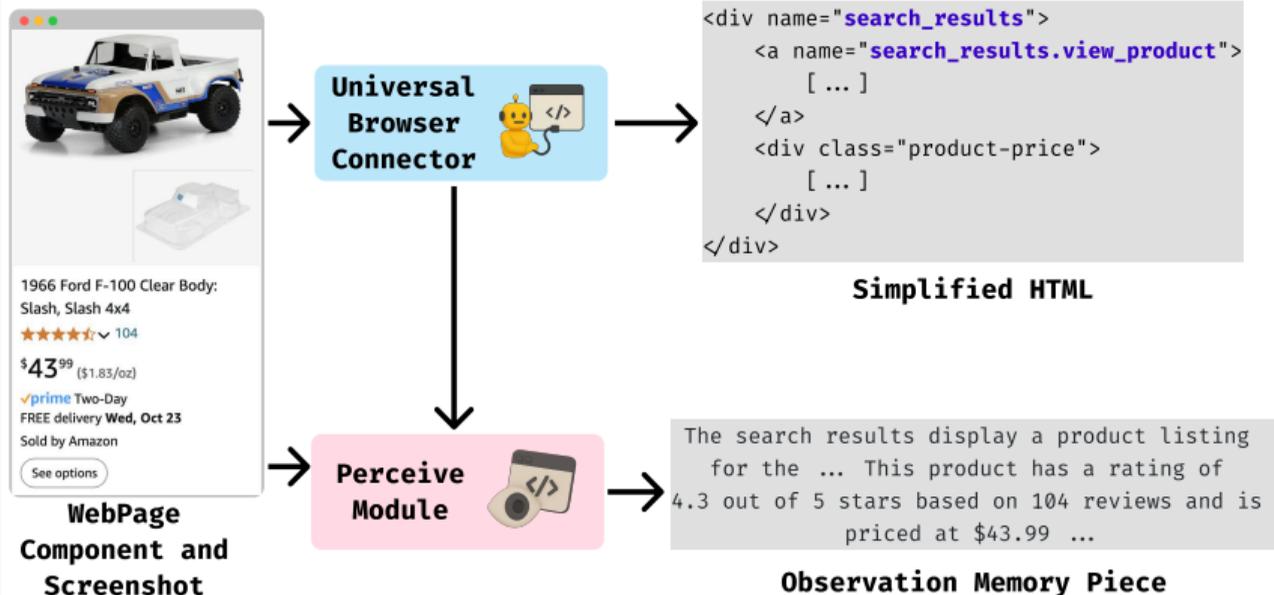


Figure 4: Universal Browser Connector

UXAgent Evaluation: can simulated agents help UX researchers?

- 16 UX researchers
- Reviewed 20 LLM simulated sessions
- Task: buy a meat substitute (highest rating, \$100–200), stop at checkout
- Flow (~30 min):
 - review draft protocol
 - inspect simulated traces/screenshots/reasoning
 - optional agent interview
 - revise protocol and feature ideas
 - survey

- Usability $M=1.63$ ($SD=0.50$); Helpfulness $M=1.38$ ($SD=0.72$)
- Trust $M=0.63$ ($SD=0.81$); Satisfaction with revised protocol $M=1.00$ ($SD=0.89$)
- “LLM can replace humans” $M=-0.44$ ($SD=1.15$)

- 14/16 proposed further study-design changes after using UXAgent
- Edits: tasks (6), demographics (3), survey items (2), interview protocol (2), added metrics (2)
- Typical cadence: build trust → sense-making → hypothesize/verify (with interviews) → converge on revisions

- Data seen as useful/rational; some behaviors “not what real users would do”
- Trust improves with transparency (reasoning trace)
- Position: complement, not replace, human participants—best for early, low-cost iteration

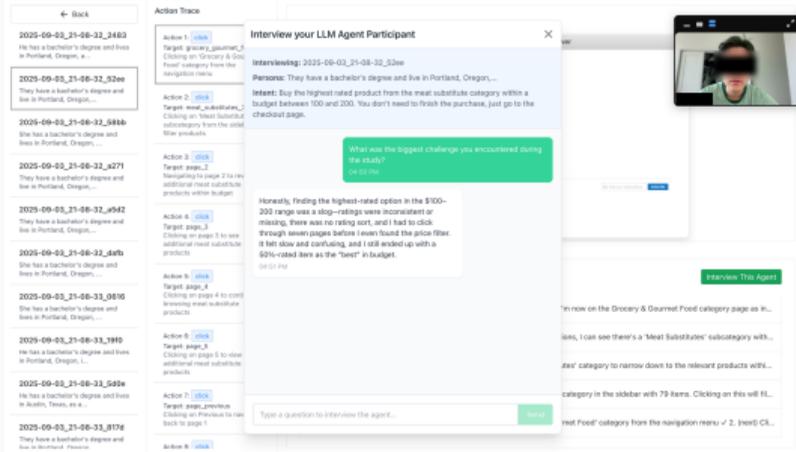


Figure 5: Study session with a UX researcher using UXAgent.

We've seen UXAgent can help UX researchers ...

We've seen UXAgent can help UX researchers ...

**Remaining question: How accurate are LLM Agents in
simulating human behavior?**

CASE STUDY 2: USE REAL ONLINE CUSTOMER BEHAVIOR DATA TO EVALUATE AND IMPROVE

CASE STUDY 2: 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.

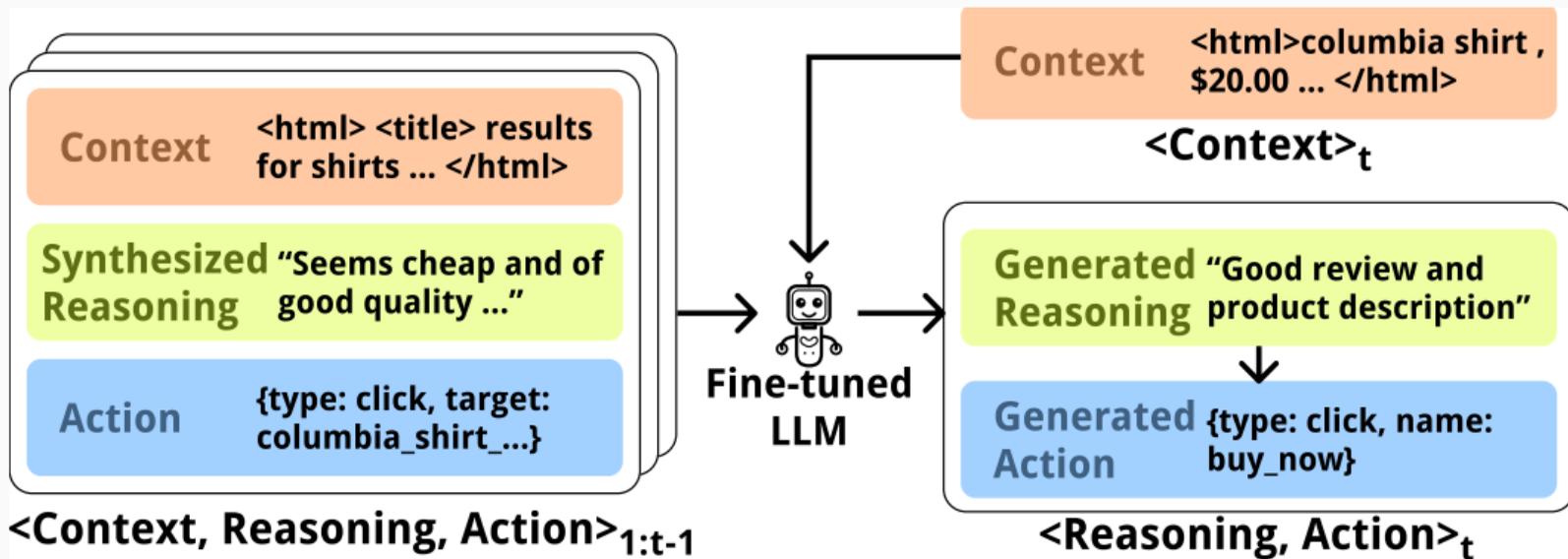


Figure 6: Overview of the next action prediction task.

- 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** (or the “observation space”) is defined as the a “simplified” HTML-based representation of the current page.
- JS and CSS are removed.
- Important structural information (Table, List, etc.) is preserved.
- LLM already understands the HTML format, no need to re-define “button” and “input” etc.

- **Context** (or the “observation space”) is defined as the a “simplified” HTML-based representation of the current page.
- JS and CSS are removed.
- Important structural information (Table, List, etc.) is preserved.
- LLM already understands the HTML format, no need to re-define “button” and “input” etc.
- Each interactable element is assigned 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.

- Reasoning traces are crucial for understanding users' action choices
- Difficult to collect; thus, they are often not available in behavioral datasets.
- Reasoning Synthesis Pipeline:
 - Record a real human customer's think-aloud shopping sessions as in-context learning examples.
 - Provide an LLM with the observation context and the corresponding action.
 - Use LLM to generate a free-text reasoning explaining the user's decision.

- To enhance LLMs' accuracy in simulating human behavior, we finetune them on the task.
 - **Input:** $\langle Context, Reasoning, Action \rangle_{1:t-1} + \langle Context \rangle_t$
 - **Output:** $\langle Reasoning, Action \rangle_t$
- Training:
 - Entire session is inputted 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 2: 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

- Baseline Models:
 - Claude
 - Llama
 - Mistral
 - DeepSeek-R1
- Fine-tuned Models:
 - Llama
 - Qwen
 - Mistral

Model	Next Action Gen.			Session Outcome		
	Acc.	%Δ	v.s. DS-R1	F1 Score	%Δ	v.s. DS-R1
Pre-Trained Models						
DeepSeek-R1	11.86%	-	-	20.01%	-	-
Mistral-7B-v0.3	4.25%	-	-7.61%	11.27%	-	-8.74%
Qwen2.5-7B	4.25%	-	-7.61%	11.94%	-	-8.07%
Llama 3.2 3B	2.93%	-	-8.93%	8.60%	-	-11.41%
Claude 3.7 Sonnet	9.34%	-	-2.52%	12.81%	-	-7.20%
Fine-tuned Models						
Qwen2.5-7B	16.67%	-	4.81%	26.92%	-	6.91%
+ reasoning	17.26%	3.54%	5.40%	33.86%	25.78%	13.85%
Mistral-7B-v0.3	14.17%	-	2.31%	17.99%	-	-2.02%
+ reasoning	15.84%	11.79%	3.98%	30.12%	67.43%	10.11%
Llama-3.2-3B	9.31%	-	-2.55%	4.73%	-	-15.28%
+ reasoning	15.77%	69.39%	3.91%	33.99%	618.60%	13.98%

Table 1: Model performance.

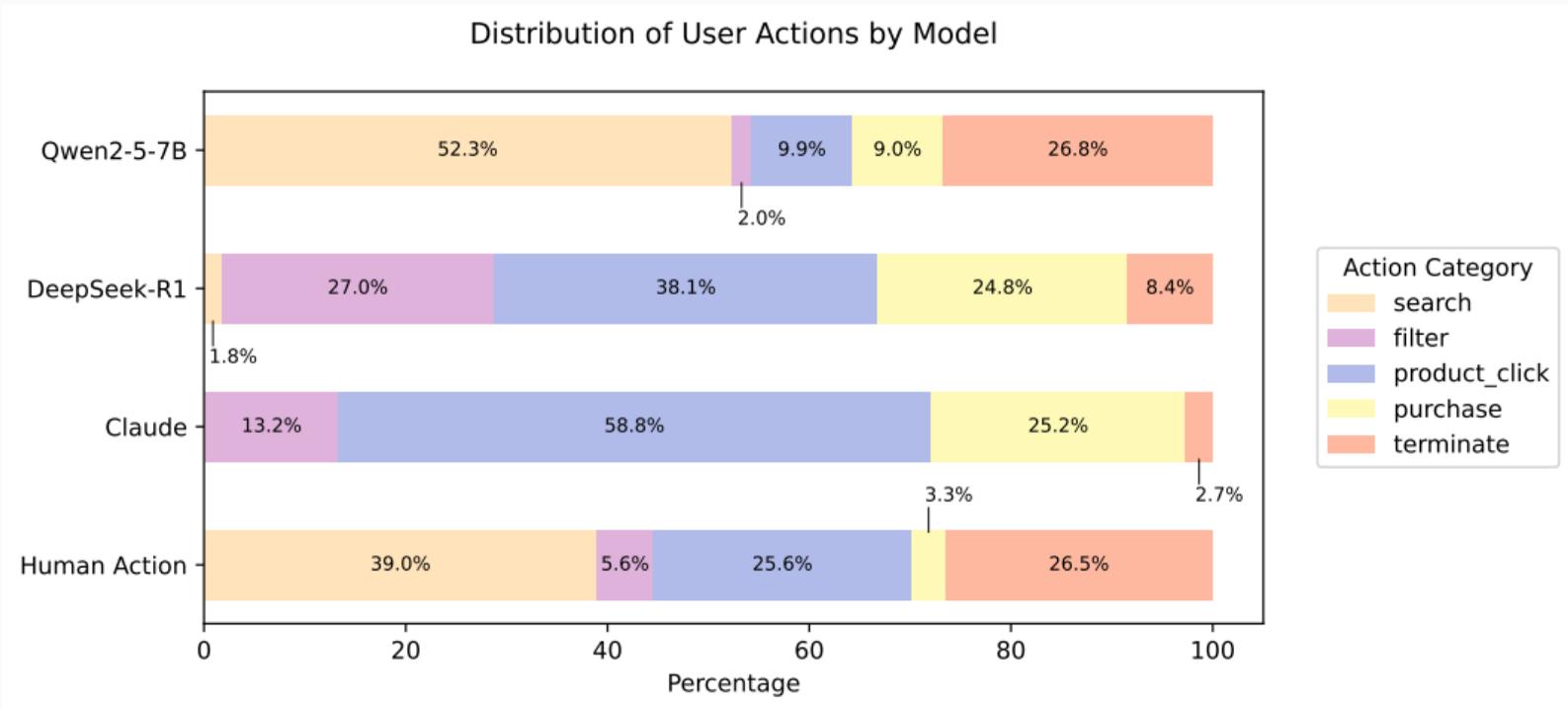


Figure 7: Distribution of the action types in the dataset.

Supervised fine-tuning improves simulation accuracy, but it remains fundamentally limited.

As an off-policy learning paradigm, SFT can only reproduce behaviors present in offline data or teacher-generated traces, rather than explore and learn new strategies.

This places an inherent ceiling on performance.

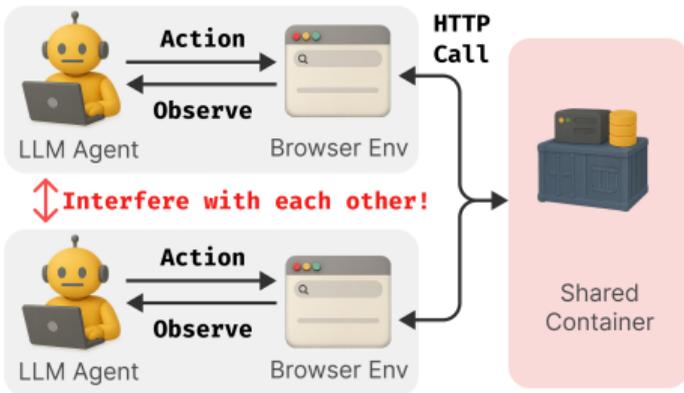
Beyond supervised fine-tuning, how can reinforcement learning further improve LLMs for human behavior simulation?

WEBSERV: A BROWSER-SERVER ENVIRONMENT FOR EFFICIENT TRAINING OF REINFORCEMENT LEARNING-BASED WEB AGENTS AT SCALE

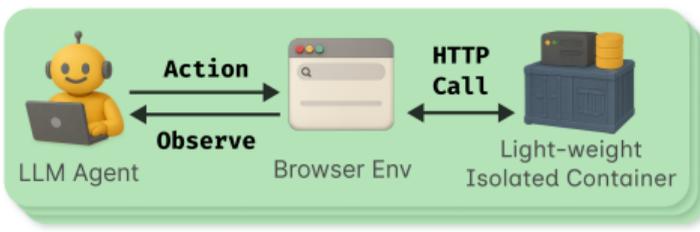
- Web Browsers are NOT RL-Ready:
 - Requires per-site configuration or rely on site features
 - Noisy context and action

- Web Servers are NOT RL-Ready
 - Expensive, slow, large image size
 - Single agent per server (lack of isolation)
 - Rollout is expensive

(a) Current Web Environment



(b) Our Full-Stack and RL-Ready WebServ



- WEBSERV + Claude 4.5: 40.0% (Shopping), 62.9% (CMS), 50.0% (GitLab) on WebArena-Lite.
- Faster spin-up: launch latency $\sim 5\times$ faster.
- Storage: $\sim 240\times$ smaller persistent storage.
- 200+ concurrent containers on a single host.

Environment	Model	Shopping	CMS	GitLab
Vanilla WebArena	GPT-4o	11.1	20.0	10.0
Vanilla WebArena	OpenAI-o3	33.3	45.7	46.7
Vanilla WebArena	Llama-3.1-8B	8.9	5.7	10.0
WEBSERV	GPT-4o	20.0	28.6	43.3
WEBSERV	OpenAI-o3	37.8	48.6	46.7
WEBSERV	Llama-3.1-8B	11.1	11.4	16.7

Table 2: Single-prompt success rate (%) on WebArena-Lite. WEBSERV lifts accuracy across proprietary and open-source models.

Metric	WEBSERV (Incus)	Naïve Docker
Launch speed	1.781 s	8.963 s
Storage	28.01 MiB	6.78 GiB
Memory	1.74 GiB	1.63 GiB

Table 3: Incus-based containers cut launch latency ($\sim 5\times$) and storage ($\sim 240\times$) at similar RAM.

- Designing: UXAgent shows how LLM agents can support UX researchers before running costly human studies.
- Evaluating: real behavioral data shows that current agents are still far from faithfully simulating humans.
- Improving: WebServ provides the infrastructure needed to train and improve web agents with reinforcement learning.

Agent A/B: Automated and Scalable A/B Testing on Live Websites with Interactive LLM Agents

Yuxuan Lu¹
Northeastern University
Boston, MA 02125
yuxuanlu@northeastern.edu

Ting-Yao Hsu¹
Computer Science and Engineering
Boston University
Boston, MA 02115

Hansu Gu
Amazon
Seattle, Washington 98144

LLM Agent Meets Agentic AI: Can LLM Agents Simulate Customers to Evaluate Agentic-AI-based Shopping Assistants?

OPeRA: A Dataset of Observation, Persona, Rationale, and Action for Evaluating LLMs on Human Online Shopping Behavior Simulation

More Samples or More Prompts? Exploring Effective Few-Shot In-Context Learning for LLMs with In-Context Sampling

Exploring Parent's Needs for Children-Centered AI to Support Preschoolers' Interactive Storytelling and Reading Activities

SHOP-R1: REWARDING LLMs TO SIMULATE HUMAN BEHAVIOR IN ONLINE SHOPPING VIA REINFORCEMENT LEARNING

Human Still Wins over LLM: An Empirical Study of Active Learning on Domain-Specific Annotation Tasks

Yuxuan Lu
Northeastern University

Bingsheng Yao
Rensselaer Polytechnic Institute

Shao Zhang
Shanghai Jiao Tong University

Yun Wang
Microsoft Research Asia

Peng Zhang
Fudan University

Tun Lu
Fudan University

Toby Ju-Jen Li
University of Notre Dame

Dakun Wang^{*}
Northeastern University

Abstract

Large Language Models (LLMs) have demonstrated considerable advances, and several claims have been made about their exceeding human performance. However, in real-world tasks, domain knowledge is often required. Low-resource learning methods like Active Learning (AL) have been proposed to tackle the cost of domain expert annotation, raising this question: **Can LLMs surpass compact models trained with expert annotations in domain-specific tasks?** In this work, we conduct an empirical experiment on four datasets from three different domains comparing SOTA LLMs with small models trained on expert annotations with AL. We found that small models outperform GPT-3.5 with a few hundreds of labeled data, and they achieve higher or similar performance with OPeRA despite that they are hundreds times smaller. Based on these findings, we posit that LLM prediction can be used as a winning method in real-world applications and human experts remain indispensable in tasks involving data annotation driven by domain-specific knowledge.

1 Introduction

Instructional-finetuned Large Language Models (Liang et al., 2022; Wei et al., 2022; Chang et al., 2022) (LLMs) have shown significant advances in their zero-shot and few-shot capabilities, i.e., in context learning (Jiang et al., 2020), on reasoning, arithmetic, and comprehension. Several recent works claim that LLMs can even outperform human crowdworkers for several data annotation tasks, such as Text Classification (Gandhi et al., 2023), Evaluation (Chang and Lee, 2023; Li et al., 2023), etc.

In reality, however, a multitude of real-world tasks necessitates the domain knowledge of human

experts in specific domains, such as doctors making diagnoses or lawyers reviewing contracts. Traditionally, such human experts have been the singular source of high quality data in their professional fields, crucial for achieving downstream goals, such as fine-tuning a language model with those high-quality annotations. However, expert annotations are frequently associated with substantial costs, limited access, and numerous challenges (Zhu et al., 2023).

In response to these challenges, the research community has put forward an array of methods to reduce the annotation cost, i.e., learning with low resources. Active learning (AL) (Elisseev, 2003; Berlin, 2002; Zhang et al., 2023) is one such popular framework to iteratively sample a few representative data, query human experts' annotations, and then train the models. AL has been demonstrated effective for various low-resource scenarios (Sharma et al., 2019; Yao et al., 2023). Thus, we ask the following question: **Do LLMs necessarily beat small models trained with limited experts' annotations in domain-specific tasks?**

We hypothesize that small language models can quickly learn domain knowledge from a limited set of labeled data with low-resource learning techniques, such as Active Learning, whereas LLMs may not perform well due to their lack of domain-specific knowledge or fine-tuning. In this work, we conduct an empirical study with Active Learning simulations on four datasets (RoBRC, Custom-NEU, Ushar, TDC, FinlaysonQA) of three types of tasks (Multiple Choice, Classification, and Question Generation) from three different domains (Biochemistry, Law, and Education). We probe the best performing state-of-the-art (SOTA) LLMs and compare them with a much smaller model finetuned with different AL techniques. Our results not only demonstrate the effectiveness of leveraging

Lin A
Paleo Arts
scholarship

Blaise
Northard
Boston, MA
lyon@psf.net

Seely
Anne
Seattle, WA
searay@psf.net



Figure 1: Architekt controlled web table

Abstract
A/B testing is central to industry product traffic, long runtimes,

Word was slow while our
Python script was fast

https://github.com/psf/psf-projects
GitHub repository

arXiv:2509.21501v1 [cs.LG] 25 Sep 2025

arXiv:2506.05606v4 [cs.CL] 24 Jul 2025

arXiv:2507.17842v2 [cs.CL] 23 Feb 2026

arXiv:2311.09825v1 [cs.CL] 16 Nov 2023

^{*}Corresponding Author: dakun@northeastern.edu

Thank you!
Questions?

