Can LLM Agents Simulate Multi-Turn Human Behavior? Evidence from Real Online Customer Behavior Data

Yuxuan Lu[1,2], Jing Huang[1], Yan Han[1], Bingsheng Yao[2], Sisong Bei[1], Jiri Gesi[1], Yaochen Xie[1], Zheshen (Jessie) Wang[1], Qi He[1], Dakuo Wang[1,2]

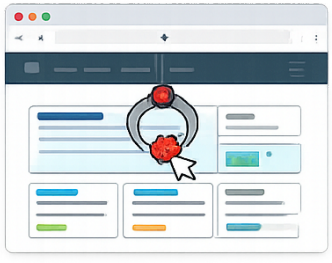
[1]Amazon.com, Inc., [2]Northeastern University

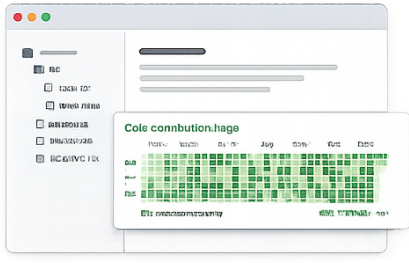
Dec 1, 2025

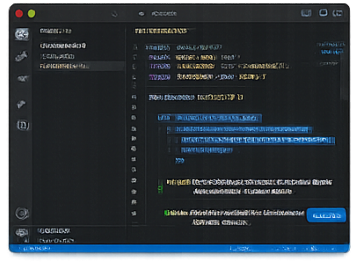
OpenClaw

opencode

Claude Code







Search homepage ⌘P

Copy path

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

Motivation

Northeastern University

- Large Language Models (LLMs) have enabled the simulation of "believable" human behavior[1].
- Many application areas have emerged:
  - Social Science Studies[2]
  - UX Studies[3]
  - A/B Testing Studies[4]

---

[1] Joon Sung Park et al. "Generative Agents: Interactive Simulacra of Human Behavior". In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. UIST '23. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 1–22.

[2] Joon Sung Park et al. *Generative Agent Simulations of 1,000 People*. Nov. 2024. arXiv: 2411.10109 [cs].

[3] Yuxuan Lu et al. *UXAgent: A System for Simulating Usability Testing of Web Design with LLM Agents*. Apr. 2025. arXiv: 2504.09407 [cs].

[4] Dakuo Wang et al. *AgentA/B: Automated and Scalable Web A/BTesting with Interactive LLM Agents*. Apr. 2025. arXiv: 2504.09723 [cs].

Motivation

Northeastern University

- However, current systems are primarily optimized for and evaluated by their "believability":
  - *"how much people feel it is like a human"*
- rather than their "accuracy":
  - *"how much it acts like a human"*

[5] Shunyu Yao et al. *ReAct: Synergizing Reasoning and Acting in Language Models*. Mar. 2023. arXiv: 2210.03629 [cs].

Motivation

Northeastern University

- However, current systems are primarily optimized for and evaluated by their "believability":
  - *"how much people feel it is like a human"*
- rather than their "accuracy":
  - *"how much it acts like a human"*
- Some work evaluates the final outcomes of tasks (e.g., item purchases)[5]

[5] Shunyu Yao et al. *ReAct: Synergizing Reasoning and Acting in Language Models*. Mar. 2023. arXiv: 2210.03629 [cs].

Motivation

Northeastern University

- However, current systems are primarily optimized for and evaluated by their "believability":
  - *"how much people feel it is like a human"*
- rather than their "accuracy":
  - *"how much it acts like a human"*
- Some work evaluates the final outcomes of tasks (e.g., item purchases)[5]
- The fidelity of intermediate actions in the sequences are not quantitatively evaluated.

[5] Shunyu Yao et al. *ReAct: Synergizing Reasoning and Acting in Language Models*. Mar. 2023. arXiv: 2210.03629 [cs].

How can we better evaluate and improve LLM Agents' action accuracy in simulating human behavior?

How can we better evaluate and improve LLM Agents' action accuracy in simulating human *shopping* behavior?

Task & Method

Task Definition

- 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.

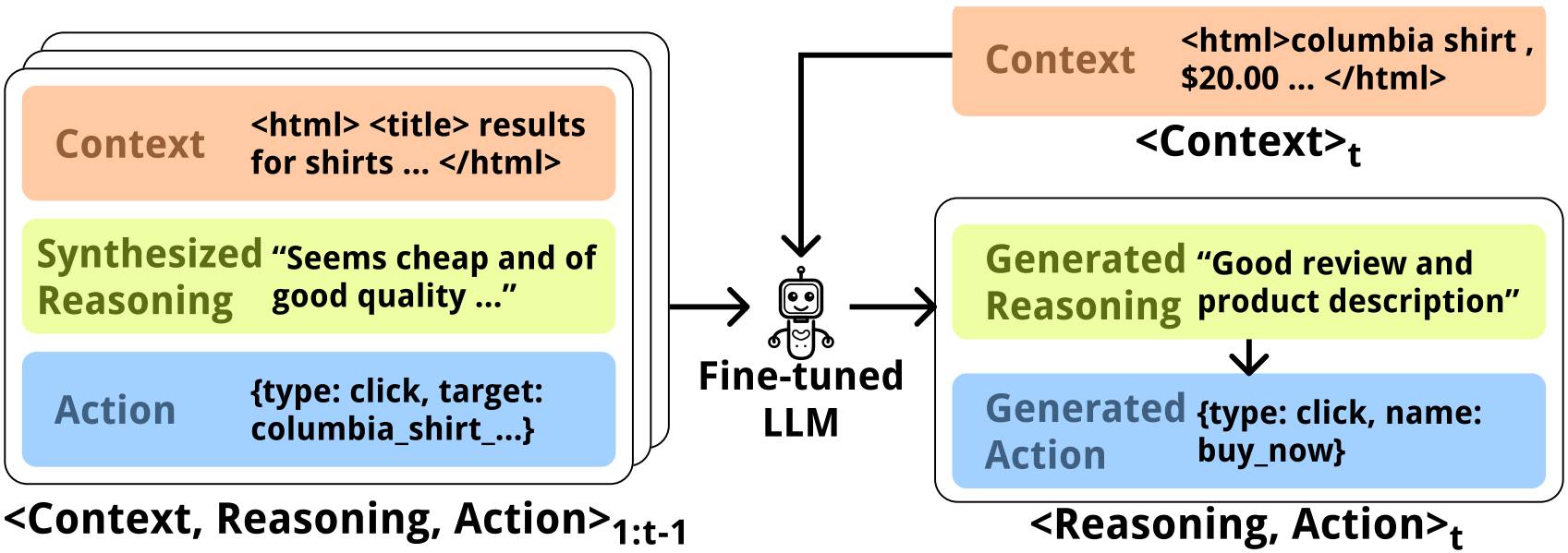


Figure 1: Overview of the next action prediction task.

Dataset

Northeastern University

- 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

- **Context** (or the "observation space") is defined as the a "simplified" HTML-based representation of the current page.

Action

- **Action** is defined as the next raw browser action conducted by the user.
  - Generalizable to other domains beyond online shopping.
- `click`
- `type_and_submit`
- `terminate`

Reasoning

Northeastern University

- **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.
- Generated by another teacher LLM

Finetuning LLMs

- 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:** $\langle Reasoning, Action \rangle_t$

Evaluation and Experiments

Evaluation Metrics

- Two tasks:
- 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.

Evaluation Metrics

- Two tasks:
- 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

Northeastern University

Model	Action Gen. (Acc.)	Outcome (F1)
Llama 3.1 70B	8.19%	12.69%
Claude 3.5 Sonnet	9.72%	15.91%
Claude 3.7 Sonnet	9.34%	12.81%
DeepSeek-R1	**11.86%**	**20.01%**
Qwen2.5-7B	4.25%	11.94%
Mistral-7B-v0.3	4.25%	11.27%
Llama 3.2 3B	2.93%	8.60%
Qwen2.5-7B SFT	**17.26%**	33.86%
Mistral-7Bv0.3 SFT	15.84%	30.12%
Llama-3.2-3B SFT	15.77%	**33.99%**

Table 1: Model performance.

Ablation Study

Northeastern University

- To evaluate the impact of training model with synthesized reasoning trace, we conduct an ablation study **to remove the reasoning trace** from the training data.

Model		Action Gen. (Acc.)	Outcome (F1)
Qwen2.5-7B SFT		**17.26%**	33.86%
	w/o reasoning	16.67%	26.92%
Mistral-7Bv0.3 SFT		15.84%	30.12%
	w/o reasoning	14.17%	17.99%
Llama-3.2-3B SFT		15.77%	**33.99%**
	w/o reasoning	9.31%	4.73%

Table 2: Ablation study result.

Conclusion

Conclusion

- We present the first quantative, process-centric evaluation of LLMs for simulating human behavior in online shopping.
- State-of-the-art models cannot simulate human behavior accurately, i.e. *Prompting is not all-you-need!*
- Fine-tuning with reasoning traces significantly improves the accuracy of LLMs in simulating human behavior.

Thank you!