Prompting is Not All You Need! Evaluating LLM Agent Simulation Methodologies with Real-World 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

June 18, 2025

Motivation

- 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

- 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

- 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

- 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.

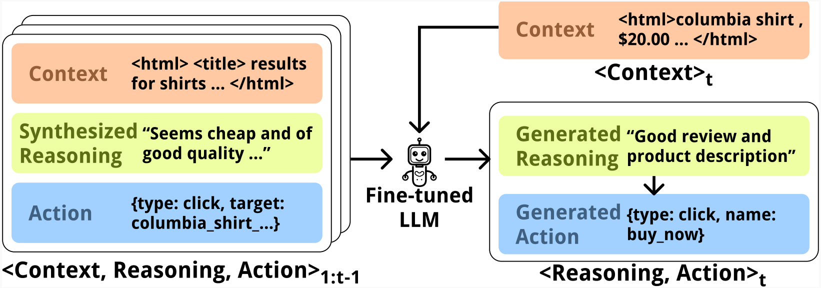

Figure 1: Overview of the next action prediction task.

Dataset

- 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.
- 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

- **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

- **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.

Rationale Synthesis

- 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.

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

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

Models

- Baseline Models:
  - Claude
  - Llama
  - Mistral
  - DeepSeek-R!
- Fine-tuned Models:
  - Llama
  - Qwen
  - Mistral

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.

Results – Distribution

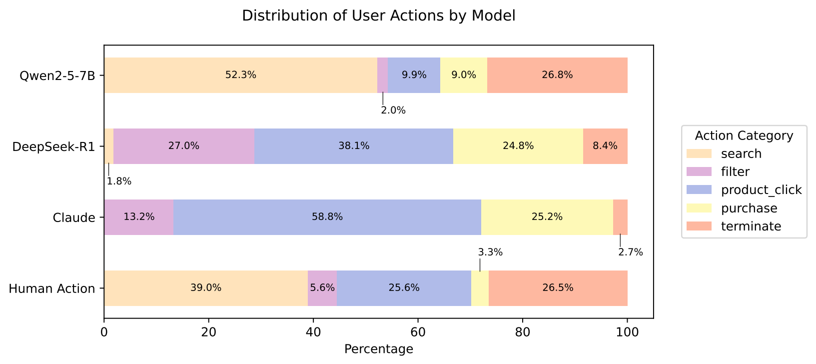

Figure 2: Distribution of the action types in the dataset.

Ablation Study

- 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.

Error Analysis

- We analyze the errors made by the models: Claude and Qwen 2.5 7B.
- Error types[6]:
  - Didn't terminate
  - Didn't click
  - Didn't search
  - Searched wrong keyword
  - Clicked wrong button

[6] Illegal actions generated by models are excluded from this analysis.

Error Analysis

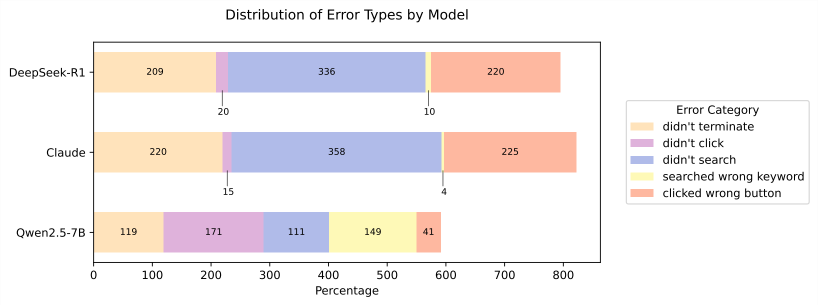

Figure 3: Error analysis.

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.

Questions? [todo] add QR Code / final conclusions