Towards Real-World Human Behavior Simulation: Benchmarking Large Language Models on Long-Horizon, Cross-Scenario, Heterogeneous Behavior Traces

Jiawei Chen[1,2]*, Ruoxi Xu[1,2]*, Boxi Cao[1], Ruotong Pan[3], Yunfei Zhang[3], Yifei Hu[3], Yong Du[3], Tingting Gao[3], Han Li[3], Yaojie Lu[1], Yingfei Sun[2], Xianpei Han[1], Le Sun[1], Xiangyu Wu[3], Hongyu Lin[1]

[1]Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences [2]University of Chinese Academy of Sciences
[3]Kuaishou Technology



Introduction

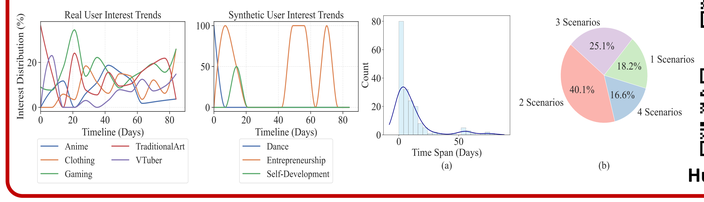
Can LLMs faithfully simulate real human behaviors?

Existing user-simulation benchmarks typically focus on isolated scenarios, narrow action spaces, or synthetic traces. However, real human behavior unfolds over long horizons, spans multiple scenarios, and combines heterogeneous actions. **OmniBehavior** fills this gap by introducing the **first user-simulation bmk** built from **real-world interaction logs** collected from Kuaishou.

Homepage

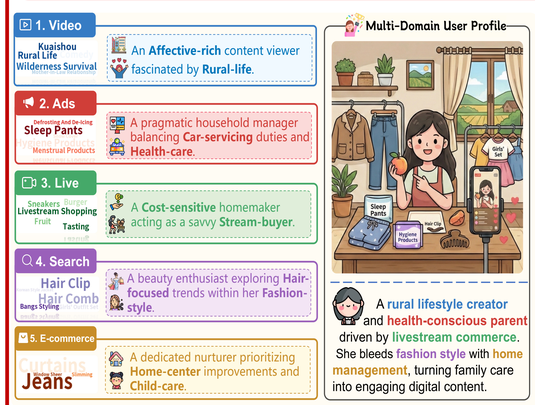
GitHub

HuggingFace

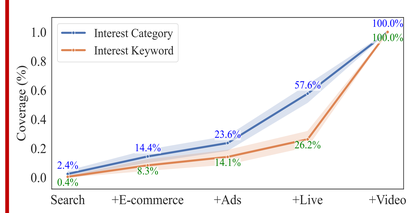


User Behavior Analysis

A. Multi-scenarios enable more comprehensive user modeling



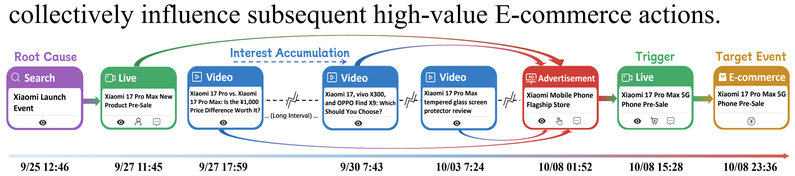
Different scenarios reveal **complementary aspects of the same user**, from content preferences and lifestyle interests to shopping intentions, consumption patterns and family needs. **No single scenario can capture the complete user profile infos.**



Search alone captures only 2.4% of interest categories and 0.4% of interest keywords observed across all scenarios. Progressively integrating **E-commerce, Ads, Live Streaming, and Video Browsing** boosts both coverage metrics to 100% across all five scenarios from Kuaishou.

B. Decisions follow long-horizon, cross-scenario causal chains

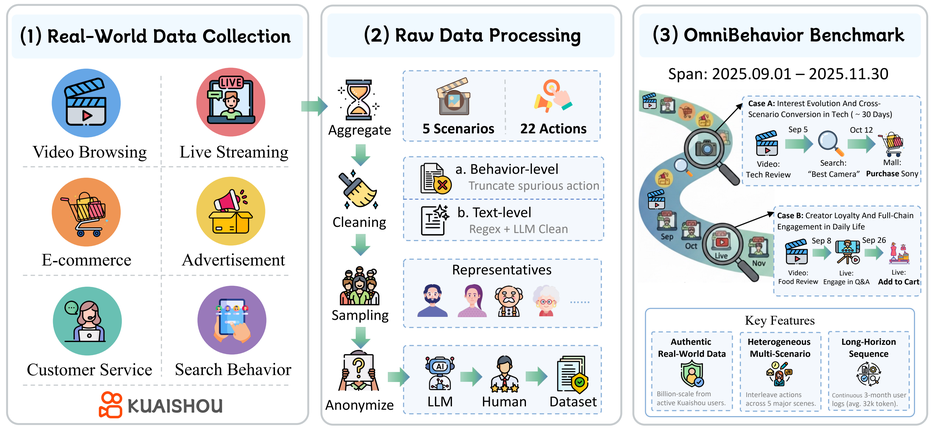
User decisions are shaped by **long-term interactions across multiple scenarios**, where earlier Search, Live-streaming, Video, and Ads collectively influence subsequent high-value E-commerce actions.



Framework

From Real-World Interactions to Predictive User Simulation

We construct **OmniBehavior** from authentic Kuaishou interaction logs across six data sources. Through a pipeline of aggregation, cleaning, sampling, and anonymization, we build the first user simulation benchmark spanning **3 months** of real-world behavior, **covering 5 scenarios and 22 action types** with long-horizon, cross-scenario user trajectories.



LLM Evaluation

We evaluate 11 representative LLMs on 6,000 prediction tasks covering **binary** (e.g., like), **continuous** (watch duration), and **textual** (dialogue) behaviors.

Model	Video		Live	Ads	E-commerce		Overall Score
	Binary	Continuous	Binary	Binary	Binary	Textual	
Human baseline							
Human*	50.21	68.83	46.79	54.76	41.46	74.63	56.11
Closed-source							
Claude-Opus-4.5	**33.05**	64.19	**31.70**	**51.16**	29.98	**57.21**	**44.55**
Claude-Sonnet-4.5	18.85	**65.95**	25.00	42.77	**36.13**	54.26	40.49
Claude-Haiku-4.5	22.84	63.26	26.11	30.00	26.37	50.29	36.48
Claude-Sonnet-4	25.29	64.62	28.86	36.81	16.50	49.13	36.87
Gemini-3-Flash	22.09	53.79	25.61	24.64	19.65	49.80	32.60
GPT-5.2	31.54	65.01	28.63	33.60	29.32	46.29	39.07
GPT-4o	27.88	62.75	28.15	25.24	28.66	44.92	36.27
Open-source							
GLM-4.7	26.86	64.43	28.97	40.34	32.90	55.25	41.46
DeepSeek-V3	21.42	63.98	27.92	25.74	33.31	52.13	37.42
Kimi-K2-Instruct	23.30	64.80	28.60	31.19	29.94	47.83	37.61
Qwen3-235B	18.26	62.38	23.84	23.19	19.22	45.74	32.11

Even the SOTA LLMs struggle to simulate real-world user behaviors.

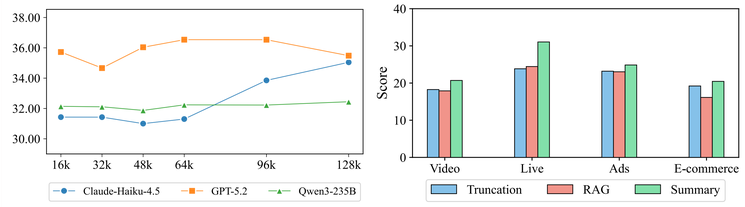
Conclusion

- ***OmniBehavior***: The first predictive user simulation benchmark **built from real-world logs**.
- ***Analysis Insights***: Long-term, cross-scenario interactions **shape more accurate** user modeling.
- ***LLM Limitations***: Even SOTA LLMs **struggle to predict behavior** with extended contexts.
- ***Structural Bias***: LLMs exhibit **"Positivity-and-Average bias"**, overestimating engagement, homogenizing users, and overlooking long-tail behaviors.

Long-context & Memory Effects

A. Longer contexts yield inconsistent gains

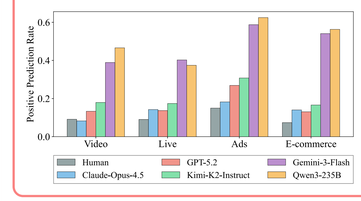
B. Memory brings only marginal gains



- Performance **does not consistently improve** from 16K to 128K
- RAG and Summarization **yield only marginal performance gains**
- Advances in **external memory**, **user modeling** may help LLMs capture behavioral signals and improve user simulation ability

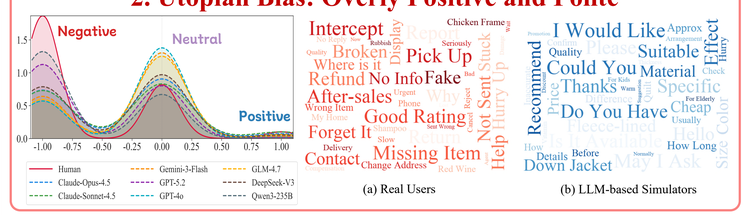
Structural Bias of LLM Simulators

1. Hyper-activity: Overestimate User Engagement

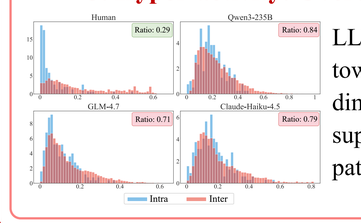


LLM simulators tend to **overestimate** positive user actions and under-model inactivity or implicit rejection, making simulated users appear more active and polite than real ones.

2. Utopian Bias: Overly Positive and Polite



3. Hyper-activity: Overestimate User Engagement



LLM-based simulated users converge towards **an average persona**, and diminishing inter-user diversity and suppressing long-tail behavioral patterns observed in real populations.

Contact

Jiawei Chen
LLM Post-training & Agent
chenjiawei30.github.io
jiawei.ucas@gmail.com

Personal Page

WeChat

Rednote