Multi-Agent Framework for Human User Simulation

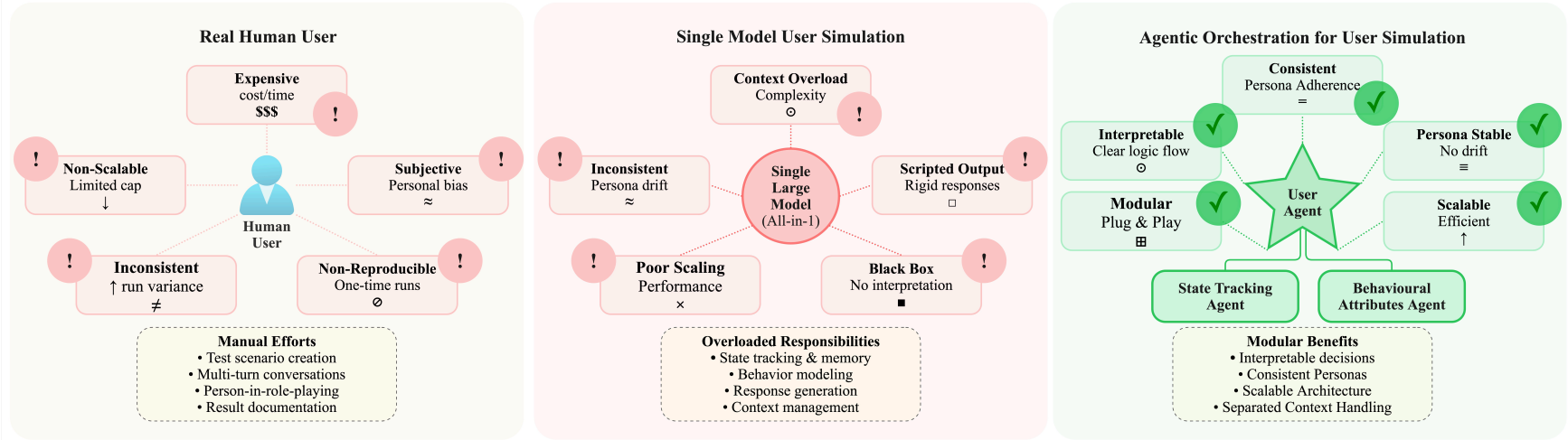
in Interactive Conversational AI Systems

A Novel Approach to Testing Conversational AI at Scale →

Read the Full Paper

9/5/2025

The Challenge



Testing conversational AI at scale is difficult

- Need realistic, diverse user interactions
- Complex behavioral patterns across personas
- Multi-turn conversation dynamics

Current approaches fall short

- Static test sets miss dynamic nature
- Human evaluators: expensive & hard to scale
- Single LLM: lacks behavioral diversity

Our Solution

Multi-Agent Orchestration

🤖

User

Primary orchestrator

Generates contextually appropriate responses based on input messages, state, and attributes

📊

State Tracking

Structured task state

Maintains current progress and target goals throughout the conversation

🎭

Message Attributes

Behavioral control

Controls mood, execution style, and exploration patterns

Agentic Setup Methodology

User Response Generation

$$r_t = f_{user}(m_t, s_t, a_t)$$

- r_t = response at turn t
- m_t = input message
- s_t = task state
- a_t = behavioral attributes

State Tracking

$$s_t = f_{stateTracking}(input\_message)$$

$$= \{\mathcal{T}_{current}, \mathcal{T}_{target}\}$$

- $\mathcal{T}_{current}$ = confirmed task items
- $\mathcal{T}_{target}$ = desired final state

Message Attributes

$$a_t = f_{msgAttrGen}(p_{bio}, s_t)$$

$$a_t = \{mood\_tone, task\_execution\_style,$$

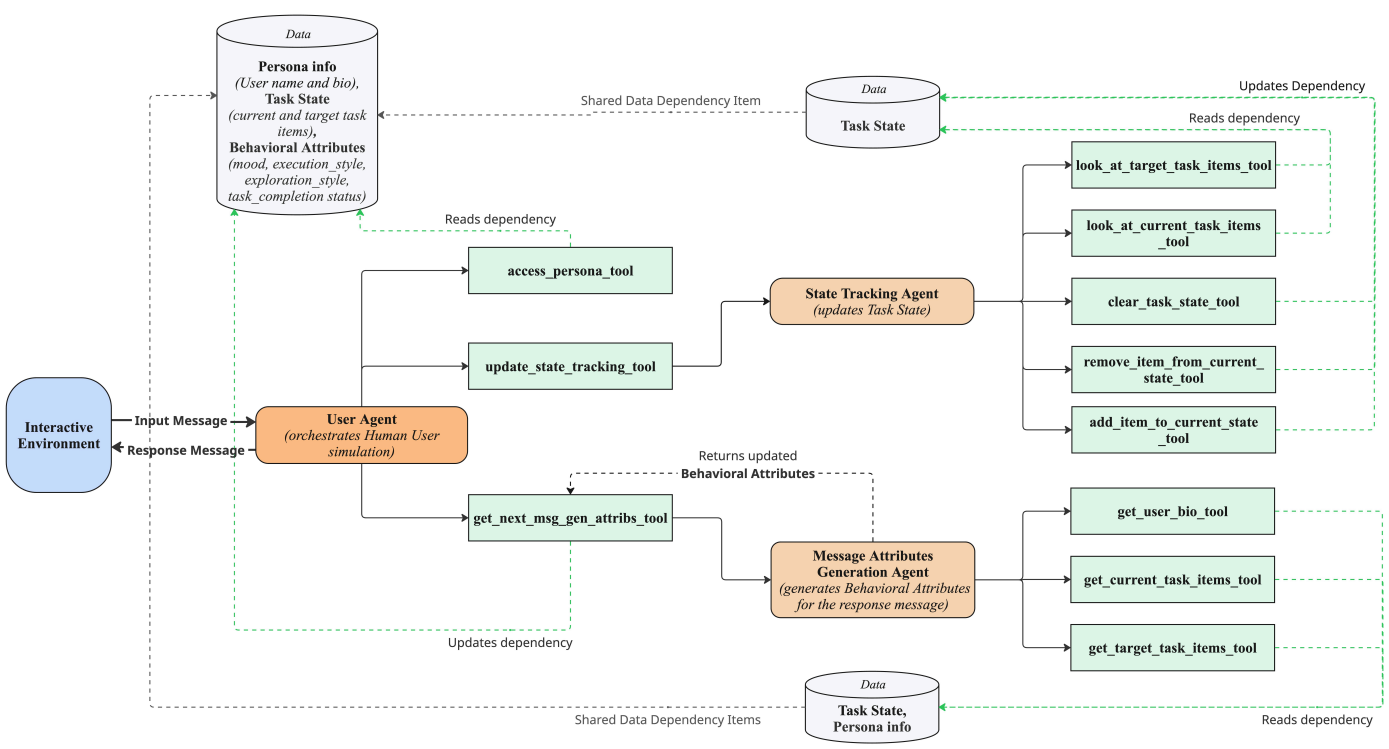
$$exploration\_style, task\_completion\_status\}$$

- p_{bio} = persona biography
- $mood\_tone$ ∈{casual, frustrated, confused, enthusiastic}
- $task\_execution\_style$ ∈{one-by-one, all-at-once}
- $exploration\_style$ ∈{explores, does-not-explore}
- $task\_completion\_status$ ∈{complete, incomplete}

Exit Gating

$$task\_completion\_status = \begin{cases} complete & \text{if } \mathcal{T}_{current} \supseteq \mathcal{T}_{target} \\ incomplete & \text{otherwise} \end{cases}$$

System Architecture



Validation Domain

Restaurant Guest Ordering

Why restaurant ordering?

✅ **Task Complexity**

- Multi-turn conversations
- Menu navigation
- Customization handling

✅ **State Management**

- Multiple items with modifiers
- Order building process
- Clarification handling

✅ **Behavioral Diversity**

- Different foodie personas
- Ordering styles
- Emotional responses

Dataset

📊 **20** diverse guest personas

🍽️ **50+** menu items with customizations

🧪 **60** test cases (3 per persona)

Implementation

- **Framework**: Pydantic AI with GPT-4o
- **Tools**: Type-safe definitions
- **Logging**: Comprehensive tracking

Ordering System

- LLM-based (GPT-4o) restaurant interface
- Natural language processing
- Menu knowledge & clarification handling
- Independent of guest simulation

Evaluation Metrics (1/2)

Persona Adherence Score (PAS)

$$PAS = \frac{1}{N}\sum_{i=1}^{N} MS_i$$

$$MS_i = \sum_{j=1}^{4} w_j \cdot C_j$$

- N = number of messages
- Equal weights $w_j = 0.25$ for:
  - Exploration style, mood tone
  - Task execution, completion status

Behavioral Variance Score (BVS)

$$TR_d = \frac{1}{M-1}\sum_{i=2}^{M} \mathbb{I}(state_i^d \neq state_{i-1}^d)$$

$$TR_{avg} = \frac{TR_{task\_execution\_style} + TR_{exploration\_style} + TR_{mood\_tone}}{3}$$

$$BVS = \begin{cases} \frac{TR_{avg}}{0.2} & \text{if } TR_{avg} \leq 0.2 \\ 1 - \frac{TR_{avg} - 0.2}{0.8} & \text{if } TR_{avg} > 0.2 \end{cases}$$

- TR_d = transition rate for behavioral dimension d
- Optimal at 20% transition rate

Task Restriction Adherence (TRA)

$$TRA = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

$$Precision = \frac{|C \cap T|}{|C|}, Recall = \frac{|C \cap T|}{|T|}$$

- T = normalized target items
- C = normalized current state items

Decision Explainability Index (DEI)

$$DEI = \begin{cases} 0 & \text{No tools} \\ \min(0.2, \frac{ED}{N} \times 0.2) & \text{Basic tools} \\ \min(0.5, \frac{ED}{N} \times 0.5) & \text{Basic + 1 agent} \\ \min(1.0, \frac{ED}{2N}) & \text{Full system} \end{cases}$$

- ED = explained decisions, N = messages

Evaluation Metrics (2/2) - Composite Score

Composite Realism & Reliability Score (CRRS)

$$CRRS = 0.25 \cdot PAS + 0.20 \cdot BVS$$

$$+0.35 \cdot TRA + 0.20 \cdot DEI$$

Unified score for overall simulation quality

Weight Distribution

- **TRA (35%)**: Primary task completion metric
- **PAS (25%)**: Persona consistency
- **BVS (20%)**: Natural behavioral variation
- **DEI (20%)**: System validation & explainability

Task completion gets highest weight as it's critical for simulation success

Metric Ranges & Interpretation

All metrics normalized to 0, 1 range:

- **PAS**: 1 = perfect persona adherence
- **BVS**: 1 = optimal variance (20% transitions)
- **TRA**: 1 = perfect F1 score for task items
- **DEI**: 1 = full explainability with tools
- **CRRS**: 1 = perfect overall simulation

Performance Targets

- **CRRS > 0.8**: Excellent simulation quality
- **CRRS 0.6-0.8**: Good simulation quality
- **CRRS < 0.6**: Needs improvement

Ablation Study Design

Configuration	Order Tracking	Message Attributes	Architecture
Config 1 (Baseline)	❌	❌	Single LLM
Config 2 (User Only)	❌	❌	Single Agent
Config 3 (User + ST)	✅	❌	Two Agents
Config 4 (User + MAG)	❌	✅	Two Agents
Config 5 (Full System)	✅	✅	Three Agents

Testing contribution of each component systematically

Results: Performance & Statistical Significance

Performance Metrics

Config	PAS	BVS	TRA	DEI	CRRS
1	0.589	0.218	0.608	0.000	0.404
2	0.585	0.485	0.582	0.200	0.487
3	0.554	0.689	**0.785**	0.498	0.651
4	**0.661**	0.000	0.602	0.432	0.462
5	0.706	**0.839**	**0.785**	**0.994**	**0.818**

🎯 **Full Multi-Agent System Achieves 102.6% Improvement**

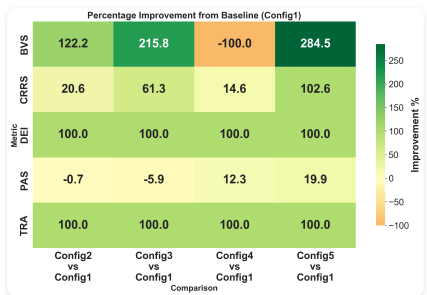
CRRS score doubles from 0.404 (baseline) to 0.818 (full system)

Statistical Significance

(Baseline vs Full System)

Metric	Δ	p-value
PAS	+20%	0.004**
BVS	+285%	<0.001***
TRA	+29%	0.005**
DEI	+100%	<0.001***
CRRS	+103%	<0.001***

** p < 0.01, *** p < 0.001



✅ All metrics significant (p < 0.01)

Key Findings

🔗 Component Synergy

- Config 3 (State only): High TRA, low PAS
- Config 4 (Attrs only): High PAS, zero BVS
- **Config 5 (Full): Best across all metrics**

> Neither sub-agent alone achieves optimal performance

🤖 Behavioral Rigidity

- Pure behavioral control without state awareness
- Results in robotic, templated interactions
- **BVS = 0 for Config 4**

> State awareness is critical for natural variance

💰 Cost-Performance Trade-off

Config	Avg Tokens	Latency (s)	CRRS
Baseline	6,618	5.08	0.404
Full System	14,789	23.16	**0.818**

2.2x tokens for 2x performance improvement

Limitations, Ethics, Future Work & Applications

Current Limitations

- 📈 High computational cost (124% more tokens)
- 🔧 Domain-specific engineering required
- 🌐 English-only validation
- 🧠 Lacks complex behaviors

Ethical Considerations

- 🔍 Maintain transparency
- ⚖️ Avoid bias perpetuation
- 🚫 No impersonation without consent
- 🛡️ Responsible deployment

Future Directions

🔄 Adaptive Evolution
Dynamic behavioral adjustment

🤹 Multi-Modal
Voice, gesture, emotional tracking

🌍 Cross-Domain
Healthcare, education, travel

⚡ Optimization
Caching, selective invocation

Applications

Testing & QA

- Automated conversation testing
- Edge case discovery
- Performance benchmarking

Broader Domains

- 🗣️ Voice assistants
- 💬 Chatbots & virtual agents
- 🏥 Healthcare interfaces
- 🛍️ E-commerce platforms
- 📚 Educational tools
- 💼 Business automation

References (1/2)

1. Ahmad et al. (2025) "Simulating User Diversity"
2. Balog & Zhai (2025) "User Simulation in the Era of Generative AI"
3. Bernard & Balog (2024) "Formal Characterization of User Simulation"
4. Castricato et al. (2024) "PERSONA: Reproducible Testbed"
5. Cheng et al. (2024) "AutoPal: Autonomous Adaptation"
6. Chu et al. (2024) "Cohesive Conversations"
7. Chu et al. (2024) "Multimodal Emotional Support"
8. Dang et al. (2025) "Multi-Agent Collaboration"
9. Davidson et al. (2023) "User Simulation with LLMs"
10. Devanathan et al. (2025) "Why Synthetic Isn’t Real Yet"
11. Feng et al. (2025) "Emotionally Intelligent Task-oriented Dialogue"
12. Ge et al. (2024) "PersonaHub: 1B Personas"
13. Hu & Ying (2025) "Unified Mind Model"
14. Hurst et al. (2024) "GPT-4o System Card"
15. Jia et al. (2024) "Leveraging LLMs for Dialogue Quality"
16. Lee et al. (2024) "OrchestraLLM: Efficient Orchestration"
17. Levi & Kadar (2025) "IntellAgent Framework"
18. Li et al. (2025) "LLM Generated Persona"
19. Liu et al. (2023) "AgentBench: Evaluating LLMs"
20. Maity & Deroy (2024) "Generative AI in Tutoring"
21. Mehri et al. (2025) "Goal Alignment in User Simulators"
22. Mo et al. (2024) "HierTOD: Hierarchical Goals"
23. Mohammadi et al. (2025) "Evaluation and Benchmarking"
24. Molchanova et al. (2025) "LLMs to Simulate Personality"
25. Niu et al. (2024) "Enhancing DST Models"
26. Park et al. (2023) "Generative Agents"
27. Park et al. (2024) "Simulations of 1,000 People"

References (2/2)

28. Park et al. (2024) "Generative Agent Simulations"
29. Park et al. (2025) "Simulating Human Behavior with AI"
30. Phy et al. (2020) "USLH Composite Metric"
31. PydanticAI (2024) "Python Agent Framework"
32. PydanticAI Docs (2024) "Documentation"
33. Rastogi et al. (2020) "Schema-Guided Dialogue"
34. Raza et al. (2024) "TRiSM for Agentic AI"
35. Saggar et al. (2025) "Score Before You Speak"
36. Shu et al. (2024) "Effective Multi-Agent Collaboration"
37. Sumers et al. (2024) "CoALA: Cognitive Architectures"
38. Sun et al. (2022) "Metaphorical User Simulators"
39. Suresh et al. (2025) "DiaSynth Framework"
40. Sutcliffe (2023) "Survey of Personality in Chatbots"
41. Tran et al. (2025) "Multi-Agent Collaboration Mechanisms"
42. Wakaki et al. (2024) "ComperDial Benchmark"
43. Wang & Chiu (2023) "Humanoid Agents Platform"
44. Wang et al. (2020) "KddRES Restaurant Dataset"
45. Wang et al. (2025) "Survey on LLM-based Agents"
46. Xiang et al. (2024) "Transformer Models for E-commerce"
47. Xie et al. (2025) "Human Simulacra Benchmark"
48. Xu et al. (2024) "Chain of Thought for DST"
49. Yi et al. (2024) "Multi-turn Dialogue Survey"
50. Yu et al. (2024) "AI Patient Simulation"
51. Zhang et al. (2025) "AgentOrchestra Framework"
52. Zhu et al. (2025) "Benchmarks & Evolutionary Evaluation"
53. Zhuge et al. (2024) "Agent-as-a-Judge"

Code & Data Availability

Implementation code, test datasets, and evaluation scripts available at:
`https://github.toasttab.com/hareeshkarthik-toast/agentic-human-guest-simulation-for-ordering`