Agentic Persona Control and Task State Tracking for Realistic User Simulation in Interactive Scenarios

Hareeshwar Karthikeyan[1]

Toast Inc., Boston, MA, USA
hareesh.karthik@toasttab.com

Abstract. Testing conversational AI systems at scale across diverse domains necessitates realistic and diverse user interactions capturing a wide array of behavioral patterns. We present a novel multi-agent framework for realistic, explainable human user simulation in interactive scenarios, using persona control and task state tracking to mirror human cognitive processes during goal-oriented conversations. Our system employs three specialized AI agents: (1) a User Agent to orchestrate the overall interaction, (2) a State Tracking Agent to maintain structured task state, and (3) a Message Attributes Generation Agent that controls conversational attributes based on task progress and assigned persona. To validate our approach, we implement and evaluate the framework for guest ordering at a restaurant with scenarios rich in task complexity, behavioral diversity, and conversational ambiguity. Through systematic ablations, we evaluate the contributory efficacy of each agentic component to overall simulation quality in terms of persona adherence, task completion accuracy, explainability, and realism. Our experiments demonstrate that the complete multi-agent system achieves superior simulation quality compared to single-LLM baselines, with significant gains across all evaluation metrics. This framework establishes a powerful environment for orchestrating agents to simulate human users with cognitive plausibility, decomposing the simulation into specialized sub-agents that reflect distinct aspects of human thought processes applicable across interactive domains.

Keywords: User Simulation · Multi-Agent Systems · AI Agents · Agentic Systems · Conversational AI · Persona Modeling · Behavioral Simulation · State Tracking · Agent Orchestration · Human-AI Interaction · Dialogue Systems · LLM Agents · Interactive Environments

1 Introduction

The rapid deployment of conversational AI systems across diverse customer-facing applications from restaurant ordering and e-commerce to healthcare consultations and customer support [32,37] has created an urgent need for comprehensive testing methodologies that can simulate realistic human user behavior [3]. Current approaches rely on static test sets or human evaluators, both presenting significant

limitations [10,2]. Static tests fail to capture the dynamic, multi-turn nature of human conversations, while human evaluation is expensive, difficult to scale, and challenging to standardize across different interaction scenarios [10,53]. Moreover, existing automated testing approaches typically lack the behavioral diversity and contextual awareness necessary to simulate realistic user interactions [2,29]. Traditional single-model approaches struggle to balance these requirements, producing either overly scripted interactions that fail to adapt, or unpredictable behaviors that compromise reliability and evaluation consistency [7,5,22,43].

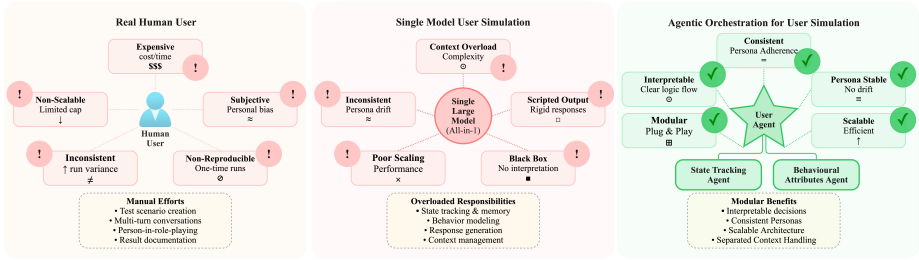

Fig. 1: **Comparison of simulation approaches: Left panel (Human User)**: Manual testing requires significant human effort including test scenario creation, multi-turn conversation management, person-in-role-playing, and result documentation, making it expensive and difficult to scale. **Center panel (Single Model System)**: Traditional automated approaches using a single model suffer from overloaded responsibilities, attempting to simultaneously handle state tracking and memory, behavior modeling, response generation, and context management, leading to inconsistent personas and poor interpretability. **Right panel (Agentic Simulation)**: Our proposed multi-agent framework distributes intelligence across specialized components, providing interpretable decisions, reproducible behavior, consistent personas, scalable architecture, and separated concerns, enabling systematic and reliable user simulation at scale. This decomposition mirrors human cognitive processes: tracking task completion progress (working memory) [37,14], deciding how to respond based on personality and context (behavioral planning) [27], and generating appropriate utterances (language production).

In this work, we propose a **multi-agent orchestration framework for human user simulation** in interactive scenarios that decomposes user behavior modeling into smaller, specialized components [9,17]. Instead of relying on a single model, the framework employs distinct agents for managing task state, generating behavioral attributes, and coordinating interactions through structured protocols. To validate our approach, we implement and evaluate the framework in restaurant ordering - a domain that reflects the complexities of human interaction through multi-turn conversations, complex state tracking, and diverse persona-driven behaviors [32]. The framework rests on three core concepts: **Task State Man-**

agement, where a State Tracking Agent maintains a structured representation of the evolving task state, enabling precise progress tracking [22,26]; **Behavioral Attribute Control**, where a Message Attributes Generation Agent dynamically determines conversational traits (mood, task execution style, exploration patterns) while preserving persona consistency [2,44]; and **Tool-mediated Coordination**, where structured protocols govern agent interactions, ensuring proper context sharing without overlap of responsibilities [33].

To our knowledge, this is the first work to explore explainable realistic human user simulation through a multi-agent architecture that combines dedicated agentic task tracking with fine-grained message generation attribute control. The unique internal environment we assess where specialized agents collaborate through structured protocols to maintain both task state coherence and persona consistency represents a novel approach in the user simulation landscape. This novelty informs our evaluation methodology, which focuses on demonstrating the framework's effectiveness through systematic ablation studies rather than direct comparisons with existing user simulation approaches that operate under fundamentally different architectural assumptions.

In summary, our work makes the following contributions:

- ✓ A novel multi-agent framework for human user simulation in interactive scenarios with specialized agents improving realism, controllability, and explainability through persona control and task state grounding
- ✓ Systematic evaluation methodology with ablation studies and standardized metrics for persona adherence, task completion accuracy, decision explainability and overall simulation quality
- ✓ Comprehensive test dataset in the restaurant ordering domain with 60 ordering test cases to validate the framework's effectiveness in complex, multi-turn conversational scenarios

2 Related Works

Human Simulation and Persona Modeling AI agents demonstrate remarkable progress in simulating human behavior. Park et al. [28,29] show generative agents replicate survey responses with 85% accuracy compared to human self-consistency, while Park et al. [27] introduce architectures combining memory, reflection, and planning for believable behavior including emergent social interactions. Persona modeling has evolved from descriptive sentences [39] to dynamic systems with internal states and emotions [44,12], though challenges remain including systematic biases [19] and personality generation difficulties [25]. Ahmad et al. [2] emphasize behavioral diversity in user simulation, while Sun et al. [37] explore metaphorical approaches. Xie et al. [46] demonstrate multi-agent cognitive mechanisms producing personified responses aligned with target characters, while Castricato et al. [5] and Ge et al. [13] procedurally generate diverse personas from demographic data. Chu et al. [7] highlight conversational coherence for maintaining persona consistency across multi-turn interactions.

Multi-Agent Orchestration and Coordination Decomposing complex tasks into specialized agents has emerged as a powerful paradigm for managing system complexity. Lee et al. [17] and Zhang et al. [50] demonstrate efficient orchestration through routing frameworks that strategically select between models, reducing computational costs by 50% while improving performance. Dang et al. [9] and Tran et al. [40] introduce dynamic orchestration with centralized coordinators trained via reinforcement learning, evolving from static to adaptive structures. Bernard and Balog [4] formalize dialogue state and action spaces for conversational systems, while Balog and Zhai [3] and Davidson et al. [10] emphasize combining LLMs with additional components to capture cognitive processes. Raza et al. [33] introduce metrics like Component Synergy Score and Tool Utilization Efficacy for quantifying collaboration quality, while Shu et al. [35] demonstrate 90% goal success rates in multi-agent collaboration, highlighting the importance of structured protocols.

Cognitive Architectures and State Management Cognitive science provides crucial insights for agent design. Sumers et al. [36] propose CoALA, drawing from symbolic AI to organize agents with modular memory components and structured action spaces mirroring human cognitive processes. Hu and Ying [14] present architectures based on Global Workspace Theory incorporating perception, planning, reasoning, memory, and motivation components. For dialogue systems, Niu et al. [26] and Xu et al. [47] use LLM-backed agents with chain-of-thought reasoning to generate annotated dialogues for state tracking, while Levi and Kadar [18] introduce graph-based modeling for multi-turn dialogues with policy constraints. Mehri et al. [22] emphasize goal alignment ensuring state tracking remains consistent with user objectives. These architectures emphasize separation between working memory (state tracking) and behavioral planning (motivation systems), validating specialized agent approaches for complex dialogue domains [32,48,23].

Synthetic Data Generation and Evaluation Agent-based systems require sophisticated evaluation methodologies beyond traditional metrics. Zhuge et al. [53] show Agent-as-a-Judge achieves 90% alignment with human consensus while reducing evaluation costs by 97%, dramatically outperforming LLM-as-a-Judge approaches. For synthetic data generation, Suresh et al. [38] use Chain of Thought reasoning to generate dialogues achieving 90.48% of in-domain data performance, while Devanathan et al. [11] introduce 18 linguistically grounded metrics revealing deficits in sentiment and behavioral realism. Evaluation frameworks must address task completion, output quality, consistency, and robustness [24], with Zhu et al. [51,52] emphasizing preventing trivial shortcuts and ensuring agents genuinely leverage persona and state understanding. Wang et al. [43] identify critical limitations in role-playing, alignment, and knowledge boundaries that multi-agent approaches can address.

3 Methodology

Our methodology employs a three-agent architecture comprising a User Agent, State Tracking Agent, and Message Attributes Generation Agent (subsections 3.1-3.3) that collaborate through structured protocols and strict behavioral rules. The system operates under defined constraints and conversation rules (subsection 3.4) to ensure reliable simulation with persona adherence and task completion accuracy.

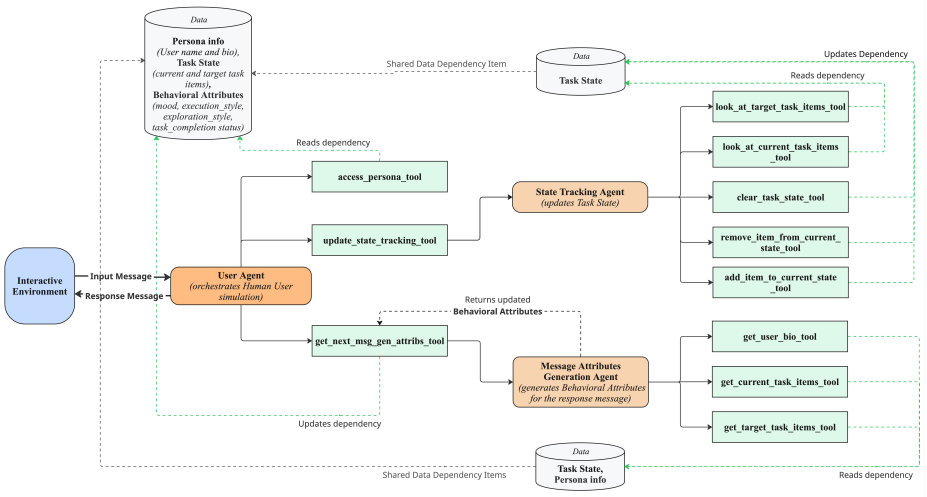

Fig. 2: **Multi-agent architecture for human user simulation** showing the three-agent framework: (1) **User Agent** serves as the primary orchestrator that generates simulated user responses by receiving input messages and invoking the two sub-agents in sequence, (2) **State Tracking Agent** maintains structured task state representation by tracking current confirmed items against target goals, and (3) **Message Attributes Generation Agent** determines behavioral characteristics (mood, execution style, exploration patterns) based on persona biography and current state.

3.1 User Agent

The User Agent serves as the primary orchestrator responsible for generating simulated user responses in the conversation [50,9]. It receives the input messages and generates contextually appropriate responses to achieve task completion, using tool calls to fetch persona info and invoke the State Tracking and Message Attributes Generation agents as needed.

The user agent's response is generated as:

$$r_t = f_{user}(m_t, s_t, a_t) \tag{1}$$

where r_t is the response at turn t, m_t is the input message, s_t is the task state from the State Tracking Agent,and a_t is the behavioral attributes from the Message Attributes Generation Agent.

3.2 State Tracking Agent

The State Tracking Agent maintains a structured representation of the current state by parsing input messages to identify task items confirmed towards achieving the target state [47,22]. The agent maintains two critical data structures:

- $\mathcal{T}_{current}$: A list of confirmed task items towards achieving the target task state
- $\mathcal{T}_{target}$: The desired final task state that the user aims to achieve.

This agent uses its tools to add, remove or clear task items in the state. It updates the task state at turn t, as:

$$s_t = f_{stateTracking}(input\_message) = \{\mathcal{T}_{current}, \mathcal{T}_{target}\} \tag{2}$$

3.3 Message Attributes Generation Agent

The Message Attributes Generation Agent determines the behavioral characteristics for each user response [12,44], while using its tool to access the persona biography and the current task state. It outputs a structured set of behavioral attributes, a_t:

$$a_t = \{mood\_tone, task\_execution\_style, exploration\_style, task\_completion\_status\} \tag{3}$$

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

This agent's decisions are conditioned as:

$$a_t = f_{msgAttrGen}(p_{bio}, s_t) \tag{4}$$

where p_{bio} is the persona biography, and s_t is the current task state.

3.4 Protocol

Baseline instructions Each agent operates with specialized system instructions that define its role and constraints. The User Agent receives instructions to maintain persona consistency while working toward task completion. The State Tracking Agent focuses solely on accurate state extraction from input messages. The Message Attributes Generation Agent balances persona traits with appropriate behavioral variation.

Critical constraints To ensure reliable simulation, our agent instructions enforce critical constraints covering [33]

1. **Tool Invocation State**: The User Agent must invoke sub-agents in a specific sequence (State Tracking → Message Attributes Generation) [40]
2. **State Consistency**: State updates must be monotonic (task completion items are only added or removed, never implicitly modified, while keeping the execution within the bounds of $\mathcal{T}_{target}$)
3. **Persona Boundaries**: Behavioral attributes must remain within persona-appropriate ranges

Conversation rules The simulation follows structured conversation rules that govern the interaction flow: beginning with initial greeting and state intent expression, proceeding through progressive state building guided by the *task_execution_style* attribute, handling clarification requests from the input, confirming state details before completion, and concluding with appropriate conversation closure once the stateing process is finished.

Exit gating The simulation terminates when the Message Attributes Generation Agent determines state completion ($task\_completion\_status = true$) [50]. This decision is based on:

$$task\_completion\_status = \begin{cases} true & \text{if } \mathcal{T}_{current} \supseteq \mathcal{T}_{target} \\ false & \text{otherwise} \end{cases} \tag{5}$$

4 Experiments

4.1 Experimental Setup

To validate our human user simulation framework, we implement and evaluate it in the domain of restaurant guest ordering. This domain presents an ideal testbed due to key characteristics aligning with human interaction complexities [42,48]. Restaurant ordering involves task complexity and ambiguity, requiring multi-turn conversations where guests navigate menu options, specify customizations, and handle clarifications [23]. The ordering process incorporates complex state tracking as guests build orders with multiple items, each having various modifiers and customization options maintained throughout the conversation. The domain captures behavioral diversity through different foodie personas exhibiting distinct ordering styles, from methodical menu explorers to decisive quick-deciders, and varying emotional responses such as frustration with overwhelming choices or confusion about menu descriptions [13]. These persona-driven mood complexities, combined with the structured yet flexible ordering task, provide an excellent environment for testing our framework's ability to balance persona consistency, task completion accuracy, and realistic behavioral variation.

Datasets

- **Personas**: 20 diverse restaurant guest personas with distinct personality traits, communication styles, and behavioral preferences [5,2]
- **Menu**: A comprehensive restaurant menu containing 50+ items across categories with various customization options and modifiers
- **Order Test Cases**: 60 test cases generated by pairing each persona with 3 different target orders of varying complexity (simple, medium, and complex orders with increasing customization levels)

Ordering System We evaluate our guest simulation by making it interact with an LLM-based ordering system (GPT-4o [15,20,35]) configured with restaurant-specific instructions and menu knowledge. The ordering system greets customers, processes natural language order requests, clarifies ambiguous requests and suggests menu items, confirms order details and handles modifications, and completes transactions with order summaries, mimicking a real restaurant environment. The ordering system operates independently of our guest simulation, receiving only the conversation history and generating responses without knowledge of the testcase information or guest system's internal state.

Agentic Simulation Implementation Our implementation leverages Pydantic AI [31] with GPT-4o for structured multi-agent development with type-safe tool definitions and dynamic data dependency injection. The multi-agent implementation consists of the following key components: a **Main User Agent** that orchestrates the Guest Agent with sub-agent access tools to fetch persona information and invoke sub-agents to update order state and generate behavioral attributes for the next message in the ordering conversation; **Sub Agents** including the Order Tracking Agent and Message Attributes Generation Agent with specialized tools to update order state and generate behavioral attributes as needed by the Guest Agent; **Data Models** comprising Pydantic Basemodel classes to hold data dependencies [1] defining the order state and behavioral attribute data objects; and **Conversation Management** featuring turn-limited interactions with repetition detection, tracking for tool calls, latencies, token usage, and structured conversation logging for comprehensive evaluation. This implementation was run on a local machine with a 10-core Apple-M1-Max CPU with 32GB of RAM and 3.2GHz of clock speed.

4.2 Simulation Evaluation Metrics

We evaluate our multi-agent framework using five quantitative metrics designed to capture different aspects of simulation quality.

Persona Adherence Score (PAS) Measures how well the user maintains their assigned persona throughout the conversation [34,41]. For each user message i in a conversation with N messages.

$$PAS = \frac{1}{N} \sum_{i=1}^{N} MS_i \tag{6}$$

where the message score MS_i is computed as:

$$MS_i = \sum_{j=1}^{4} w_j \cdot C_j \tag{7}$$

with equal weights $w_j = 0.25$ for each component:

- C_1: Exploration style match (explores vs. does not explore)
- C_2: Mood tone alignment (casual, frustrated, confused, enthusiastic)
- C_3: Task execution style match (one-by-one vs. all-at-once)
- C_4: Task completion status agreement

Each component $C_j \in \{0, 1\}$ is computed based on exact match with expected persona attributes.

Behavioral Variance Score (BVS) Captures realistic fluctuations in behavior to ensure natural human-like variations. For each behavioral dimension $d \in \{task\_execution\_style, exploration\_style, mood\_tone\}$:

$$TR_d = \frac{1}{M-1} \sum_{i=2}^{M} \mathbb{I}(state_i^d \neq state_{i-1}^d) \tag{8}$$

where M is the number of behavioral states. The average transition rate is:

$$TR_{avg} = \frac{TR_{task\_execution\_style} + TR_{exploration\_style} + TR_{mood\_tone}}{3} \tag{9}$$

BVS uses a piecewise linear scoring function peaking at 20% transition rate since humans typically exhibit moderate variance [28] :

$$BVS = \begin{cases} \frac{TR_{avg}}{0.2} & \text{if } TR_{avg} \leq 0.2 \\ 1 - \frac{TR_{avg} - 0.2}{0.8} & \text{if } TR_{avg} > 0.2 \end{cases} \tag{10}$$

The range of BVS is [0, 1] where 1 is the best possible score. So we can expect robotic (too static) patterns having lower BVS scores while realistic patterns would have higher BVS scores.

Task Restriction Adherence (TRA) Evaluates accuracy in achieving the target state using F1-score with normalized task item matching [16]. Task items are normalized by:

$$normalize(item) = lowercase(remove\_special(remove\_filler(item))) \tag{11}$$

With T = normalized target items and C = normalized current state items,

$$Precision = \frac{|C \cap T|}{|C|}, \quad Recall = \frac{|C \cap T|}{|T|} \tag{12}$$

$$TRA = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{13}$$

Decision Explainability Index (DEI) Quantifies the traceability of the agentic system's internal decisions with tool usage results [33].

$$DEI = \begin{cases} 0 & \text{No tools (no explainability)} \\ \min(0.2, \frac{ED}{N} \times 0.2) & \text{Basic tools (about 20\% explainability)} \\ \min(0.5, \frac{ED}{N} \times 0.5) & \text{Basic tools + 1 subAgent (about 50\% explainability)} \\ \min(1.0, \frac{ED}{2N}) & \text{Full system (100\% explainability)} \end{cases} \tag{14}$$

where ED is the count of explained decisions (tool invocations) across N messages.

Composite Realism and Reliability Score (CRRS) Provides a unified score for overall user simulation quality using universal weights [30].

$$CRRS = 0.25 \cdot PAS + 0.20 \cdot BVS + 0.35 \cdot TRA + 0.20 \cdot DEI \tag{15}$$

These weights reflect: TRA (35%) as the primary task metric, PAS (25%) for persona consistency, BVS (20%) for naturalness, and DEI (20%) for system validation. Note that DEI naturally adjusts based on the agentic system's experimental setup.

4.3 Ablations

We conduct systematic ablation studies across five experimental configurations to isolate the contribution of each component.

Config1 - Baseline LLM: Single LLM with all information (persona, target order, conversation history) provided directly in the prompt. No agentic decomposition or tool use.

Config2 - User Agent Only: Guest Agent without sub-agents. Has direct access to persona, target order and conversation history through tools but no structured state tracking or behavioral control.

Config3 - User Agent + State Tracking (ST) Agent: Guest Agent with Order Tracking sub-agent. Maintains structured order state but lacks explicit behavioral control.

Config4 - User Agent + Message Attributes Generation (MAG) Agent: Guest Agent with Message Attributes Generation sub-agent. Has behavioral control but no structured state tracking.

Config5 - Full System: Complete multi-agent system with both Order Tracking and Message Attributes Generation sub-agents.

4.4 Results

We evaluate each configuration across five quantitative metrics defined in Section 3.3. All experiments are run with 60 test cases per configuration.

Table 1: Evaluation Metrics Across All Configurations

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**

Table 2: Response Computation Costs

Config	Avg. Tokens	Avg. Latency(s)
1	6,618	5.08
2	13,505	4.56
3	24,580	36.30
4	15,763	16.88
5	14,789	23.16

Tables 1 and 2 present the complete evaluation results and the computational costs incurred per response while executing each configuration.

Table 3: Statistical Significance: Full System vs Baseline

Metric	p-value	Improve
PAS	0.0037**	+19.9%
BVS	0.0000***	+284.5%
TRA	0.0047**	+29.1%
DEI	0.0000***	+100.0%
CRRS	0.0000***	+102.6%

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

Table 3 presents the statistical significance analysis and Figure 3 visualizes the performance improvements across all configurations.

Metrics Analysis Persona Adherence Score (PAS): Config4 achieves the highest individual PAS (0.661), demonstrating the Message Attributes Agent's effectiveness for persona consistency. The full system (Config5) shows a 19.9% improvement over baseline.

Fig. 3: Performance gains from baseline

Behavioral Variance Score (BVS): Config5 achieves the most realistic behavioral variance (0.839), representing a 284.5% improvement over the static baseline. Notably, Config4 shows zero variance, indicating over-rigid behavioral control without order state awareness.

Task Restriction Adherence (TRA): Both Config3 and Config5 achieve equivalent high performance (0.785), confirming the State Tracking Agent's critical role in maintaining order accuracy.

Decision Explainability Index (DEI): Only Config5 achieves near-perfect explainability (0.994) through comprehensive tool usage traces, while simpler configurations lack decision transparency.

Composite Realism and Reliability Score (CRRS): Config5 demonstrates the highest overall performance (0.818), achieving a **102.6% simulation quality improvement** over baseline and outperforming all partial configurations.

Key Findings Component Synergy: Neither sub-agent alone (Config3, Config4) achieves optimal performance; the combination (Config5) creates synergistic effects.

Behavioral Rigidity: Pure behavioral control (Config4) without state awareness leads to robotic, templated persona-default style interactions (BVS = 0).

Cost-Performance Trade-off: The full system balances computational cost (14,789 tokens) with superior performance across all metrics.

Statistical Significance: All improvements in Config5 over baseline are statistically significant ($p < 0.01$), indicating the effectiveness of our multi-agent architecture across all metrics for simulating human user behaviour in a task oriented conversation.

5 Conclusion

We presented a multi-agent framework decomposing human user simulation into specialized components: a User Agent for orchestration, State Tracking Agent for task management, and Message Attributes Generation Agent for behavioral control. Through persona control and task state grounding, our framework enables realistic simulation in conversational domains. Validating it in restaurant guest ordering scenarios demonstrates 102.6% improvement in simulation quality over single-LLM baselines, with significant gains in persona adherence (19.9%), behavioral variance (284.5%), and task completion accuracy (29.1%). These results validate that decomposing human simulation into specialized agents with coordinated state management yields superior performance across simulation quality dimensions, establishing our multi-agent architecture as effective for realistic human simulation in interactive AI systems.

Limitations and Ethical Considerations The multi-agent architecture incurs substantial computational costs (124% more tokens, 356% higher latency), limiting resource-constrained deployment. Each domain requires specialized prompt engineering and state design. Behavioral modeling lacks complex human behaviors like indecision, social dynamics, or cultural nuances [43]. Validation is English-only, single-domain with 60 test cases, restricting cross-cultural, multilingual, and real-world insights. Systems must maintain transparency and avoid bias perpetuation through auditing [19]. While democratizing conversational AI testing, the framework shouldn't impersonate individuals without consent or manipulate users believing they interact with real humans.

Future Work and Applications Key directions include adaptive persona evolution enabling dynamic behavioral adjustment [6], multi-modal integration for voice-based prosodic features, visual gesture recognition, and emotional sentiment tracking [8], and cross-domain validation across customer support, e-commerce ordering [45], healthcare consultations [49], educational tutoring [21], financial advisory, travel booking, and technical troubleshooting. Efficiency optimization through agent caching, selective tool invocation, and smaller specialized models could address computational overhead. The core principle of decomposing human behavioral simulation into specialized agents managing persona consistency and task state provides a foundation for realistic user simulations across interactive domains. Our framework offers systematic high-quality interaction data generation for testing, evaluation, and quality assurance wherever validating human-system interactions is critical for system reliability and user experience.

Acknowledgments. Thanks to Andrew Norfleet (Principal Data Scientist at Toast) for his invaluable support during the ideation of the agentic simulation piece, and Benjamin Tang (Director of AI at Toast) for his support in pursuing the development of this work. This work was funded by Toast Inc.

Disclosure of Interests. The author is employed by Toast Inc., which funded this research. The author declares no other competing interests relevant to the content of this article.

References

1. Pydanticai documentation. `https://ai.pydantic.dev/` (2024), accessed: 2025-09-01
2. Ahmad, A., Hillmann, S., Möller, S.: Simulating User Diversity in Task-Oriented Dialogue Systems using Large Language Models. In: arXiv preprint arXiv:2502.12813 (2025)
3. Balog, K., Zhai, C.: User Simulation in the Era of Generative AI: User Modeling, Synthetic Data Generation, and System Evaluation. In: arXiv preprint arXiv:2501.04410v1 (2025)
4. Bernard, N., Balog, K.: Towards a Formal Characterization of User Simulation Objectives in Conversational Information Access. In: arXiv preprint arXiv:2406.19007v1 (2024)
5. Castricato, L., Lile, N., Rafailov, R., Fränken, J.P., Finn, C.: PERSONA: A Reproducible Testbed for Pluralistic Alignment. In: arXiv preprint arXiv:2407.17387 (2024)
6. Cheng, Y., Liu, W., Xu, K., Hou, W., Ouyang, Y., Leong, C.T., Li, W., Wu, X., Zheng, Y.: AutoPal: Autonomous Adaptation to Users for Personal AI Companionship. In: arXiv preprint arXiv:2406.13960 (2024)
7. Chu, K., Chen, Y.P., Nakayama, H.: Cohesive Conversations: Enhancing Authenticity in Multi-Agent Simulated Dialogues. In: arXiv preprint arXiv:2407.09897 (2024)
8. Chu, Y., Liao, L., Zhou, Z., Ngo, C.W., Hong, R.: Towards Multimodal Emotional Support Conversation Systems. In: arXiv preprint arXiv:2408.03650 (2024)
9. Dang, Y., Qian, C., Luo, X., et al.: Multi-Agent Collaboration via Evolving Orchestration. In: arXiv preprint arXiv:2505.19591 (2025)
10. Davidson, S., Romeo, S., Shu, R., Gung, J., Gupta, A., Mansour, S., Zhang, Y.: User Simulation with Large Language Models for Evaluating Task-Oriented Dialogue. In: arXiv preprint arXiv:2309.13233 (2023)
11. Devanathan, R., Nathan, V., Kumar, A.: Why Synthetic Isn't Real Yet: A Diagnostic Framework for Contact Center Dialogue Generation. In: arXiv preprint arXiv:2508.18210 (2025)
12. Feng, S., chin Lin, H., Lubis, N., van Niekerk, C., Heck, M., Ruppik, B., Vukovic, R., Gašić, M.: Emotionally Intelligent Task-oriented Dialogue Systems: Architecture, Representation, and Optimisation. In: arXiv preprint arXiv:2507.01594 (2025)
13. Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., Yu, D.: Scaling Synthetic Data Creation with 1,000,000,000 Personas. In: arXiv preprint arXiv:2406.20094 (2024)
14. Hu, P., Ying, X.: Unified Mind Model: Reimagining Autonomous Agents in the LLM Era. In: arXiv preprint arXiv:2503.03459 (2025)
15. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A.T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., et al.: GPT-4o System Card. In: arXiv preprint arXiv:2410.21276 (2024)

16. Jia, J., Komma, A., Leffel, T., Peng, X., Nagesh, A., Soliman, T., Galstyan, A., Kumar, A.: Leveraging LLMs for Dialogue Quality Measurement. In: arXiv preprint arXiv:2406.17304 (2024)
17. Lee, C.H., Cheng, H., Ostendorf, M.: OrchestraLLM: Efficient Orchestration of Language Models for Dialogue State Tracking. In: arXiv preprint arXiv:2311.09758 (2024)
18. Levi, E., Kadar, I.: IntellAgent: A Multi-Agent Framework for Evaluating Conversational AI Systems. In: arXiv preprint arXiv:2501.11067 (2025)
19. Li, A., Chen, H., Namkoong, H., Peng, T.: LLM Generated Persona is a Promise with a Catch. In: arXiv preprint arXiv:2503.16527v1 (2025)
20. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., Tang, J.: AgentBench: Evaluating LLMs as Agents. In: arXiv preprint arXiv:2308.03688 (2023)
21. Maity, S., Deroy, A.: Generative AI and Its Impact on Personalized Intelligent Tutoring Systems. In: arXiv preprint arXiv:2410.10650 (2024)
22. Mehri, S., Yang, X., Kim, T., Tur, G., Mehri, S., Hakkani-Tür, D.: Goal Alignment in LLM-Based User Simulators for Conversational AI. In: arXiv preprint arXiv:2507.20152 (2025)
23. Mo, L., Jiang, S., Maharaj, A., Hishamunda, B., Li, Y.: HierTOD: A Task-Oriented Dialogue System Driven by Hierarchical Goals. In: arXiv preprint arXiv:2411.07152 (2024)
24. Mohammadi, M., Li, Y., Lo, J., Yip, W.: Evaluation and Benchmarking of LLM Agents: A Survey. In: arXiv preprint arXiv:2507.21504v1 (2025)
25. Molchanova, M., Mikhailova, A., Korzanova, A., Ostyakova, L., Dolidze, A.: Exploring the Potential of Large Language Models to Simulate Personality. In: Workshop on Customizable NLP (CustomNLP4U) at EMNLP 2024 (2025)
26. Niu, C., Wang, X., Cheng, X., Song, J., Zhang, T.: Enhancing Dialogue State Tracking Models through LLM-backed User-Agents Simulation. In: arXiv preprint arXiv:2405.13037 (2024)
27. Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative Agents: Interactive Simulacra of Human Behavior. In: arXiv preprint arXiv:2304.03442 (2023)
28. Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C., Morris, M.R., Willer, R., Liang, P., Bernstein, M.S.: Generative Agent Simulations of 1,000 People. In: arXiv preprint arXiv:2411.10109 (2024)
29. Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C.J., Morris, M.R., Willer, R., Liang, P., Bernstein, M.S.: Simulating human behavior with ai agents. Policy brief, Stanford Institute for Human-Centered Artificial Intelligence (HAI) (May 2025), https://hai.stanford.edu/assets/files/hai-policy-brief-simulating-human-behavior-with-ai-agents.pdf
30. Phy, V., Zhao, Y., Aizawa, A.: Deconstruct to Reconstruct a Configurable Evaluation Metric for Open-Domain Dialogue Systems. In: Proceedings of COLING 2020 (2020)
31. Pydantic Team: Pydanticai: A python agent framework for llms. https://github.com/pydantic/pydantic-ai (2024), accessed: 2025-09-01
32. Rastogi, A., Zang, X., Sunkara, S., Gupta, R., Khaitan, P.: Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset. In: Proceedings of AAAI 2020 (2020)
33. Raza, S., Sapkota, R., Karkee, M., Emmanouilidis, C.: TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems. In: arXiv preprint arXiv:2506.04133 (2024)

34. Saggar, A., Darling, J.C., Dimitrova, V., Sarikaya, D., Hogg, D.C.: Score Before You Speak: Improving Persona Consistency in Dialogue Generation using Response Quality Scores. In: Proceedings of ECAI 2025 (2025)
35. Shu, R., Das, N., Yuan, M., Sunkara, M., Zhang, Y.: Towards Effective GenAI Multi-Agent Collaboration: Design and Evaluation for Enterprise Applications. In: arXiv preprint arXiv:2412.05449 (2024)
36. Sumers, T.R., Yao, S., Narasimhan, K., Griffiths, T.L.: Cognitive Architectures for Language Agents. In: arXiv preprint arXiv:2309.02427 (2024)
37. Sun, W., Guo, S., Zhang, S., Ren, P., Chen, Z., de Rijke, M., Ren, Z.: Metaphorical User Simulators for Evaluating Task-oriented Dialogue Systems. In: arXiv preprint arXiv:2204.00763 (2022)
38. Suresh, S.K., Mengjun, W., Pranav, T., Chng, E.S.: DiaSynth: Synthetic Dialogue Generation Framework for Low Resource Dialogue Applications. In: Proceedings of NAACL 2025 (2025)
39. Sutcliffe, R.: A Survey of Personality, Persona, and Profile in Conversational Agents and Chatbots. In: arXiv preprint arXiv:2401.00609 (2023)
40. Tran, K.T., Dao, D., Nguyen, M.D., Pham, Q.V., O'Sullivan, B., Nguyen, H.D.: Multi-Agent Collaboration Mechanisms: A Survey of LLMs. In: arXiv preprint arXiv:2501.06322 (2025)
41. Wakaki, H., Mitsufuji, Y., Maeda, Y., Nishimura, Y., Gao, S., Zhao, M., Yamada, K., Bosselut, A.: ComperDial: Commonsense Persona-grounded Dialogue Dataset and Benchmark. In: arXiv preprint arXiv:2406.11228 (2024)
42. Wang, H., Li, M., Zhou, Z., Fung, G.P.C., Wong, K.F.: KddRES: A Multi-level Knowledge-driven Dialogue Dataset for Restaurant Towards Customized Dialogue System. In: arXiv preprint arXiv:2011.08772 (2020)
43. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z.Y., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z., Wen, J.R.: A Survey on Large Language Model based Autonomous Agents. In: arXiv preprint arXiv:2308.11432 (2025)
44. Wang, Z., Chiu, Y.Y., Chiu, Y.C.: Humanoid Agents: Platform for Simulating Human-like Generative Agents. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 167–176. Association for Computational Linguistics, Singapore (2023)
45. Xiang, Y., Yu, H., Gong, Y., Huo, S., Zhu, M.: Text Understanding and Generation Using Transformer Models for Intelligent E-commerce Recommendations. In: arXiv preprint arXiv:2402.16035 (2024)
46. Xie, Q., Feng, Q., Zhang, T., et al.: Human Simulacra: Benchmarking the Personification of Large Language Models. In: Proceedings of ICLR 2025 (2025)
47. Xu, L., Peng, N., Zhou, D., Ng, S.K., Fu, J.: Chain of Thought Explanation for Dialogue State Tracking. In: arXiv preprint arXiv:2403.04656 (2024)
48. Yi, Z., Ouyang, J., Xu, Z., Liu, Y., Liao, T., Luo, H., Shen, Y.: A Survey on Recent Advances in LLM-Based Multi-turn Dialogue Systems. In: arXiv preprint arXiv:2402.18013 (2024)
49. Yu, H., Zhou, J., Li, L., et al.: Simulated patient systems are intelligent when powered by large language model-based AI agents. In: arXiv preprint arXiv:2409.18924 (2024)
50. Zhang, W., Zeng, L., Xiao, Y., et al.: AgentOrchestra: A Hierarchical Multi-Agent Framework for General-Purpose Task Solving. In: arXiv preprint arXiv:2506.12508 (2025)
51. Zhu, J., Zhu, M., Rui, R., Shan, R., Zheng, C., Chen, B., Xi, Y., Lin, J., Liu, W., Tang, R., Yu, Y., Zhang, W.: Evolutionary Perspectives on the Evaluation of LLM-Based AI Agents: A Comprehensive Survey. In: arXiv preprint arXiv:2506.11102v1 (2025)

52. Zhu, Y., Jin, T., Pruksachatkun, Y., Zhang, A., Liu, S., Cui, S., Kapoor, S., Longpre, S., Meng, K., Weiss, R., Barez, F., Gupta, R., Dhamala, J., Merizian, J., Giulianelli, M., Coppock, H., Ududec, C., Sekhon, J., Steinhardt, J., Kellerman, A., Schwettmann, S., Zaharia, M., Stoica, I., Liang, P., Kang, D.: Establishing Best Practices for Building Rigorous Agentic Benchmarks. In: arXiv preprint arXiv:2507.02825v2 (2025)
53. Zhuge, M., Zhao, C., Ashley, D.R., Wang, W., Khizbullin, D., Xiong, Y., Liu, Z., Chang, E., Krishnamoorthi, R., Tian, Y., Shi, Y., Chandra, V., Schmidhuber, J.: Agent-as-a-Judge: Evaluate Agents with Agents. In: arXiv preprint arXiv:2410.10934 (2024)