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

Hareeshwar Karthikeyan
Data Scientist
Toast Inc.
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.

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].

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

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Scaling Environments for Agents (SEA).

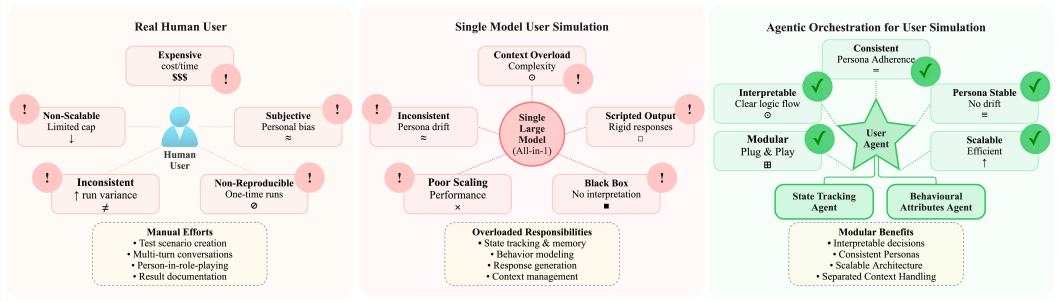

Figure 1: **Comparison of simulation approaches: Left panel (Human User)**: Manual testing requires significant human effort including test scenario creation, multi-turn conversation management, person-in-role-playing, and result documentation, making it expensive and difficult to scale. **Center panel (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).

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 Management**, 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.

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

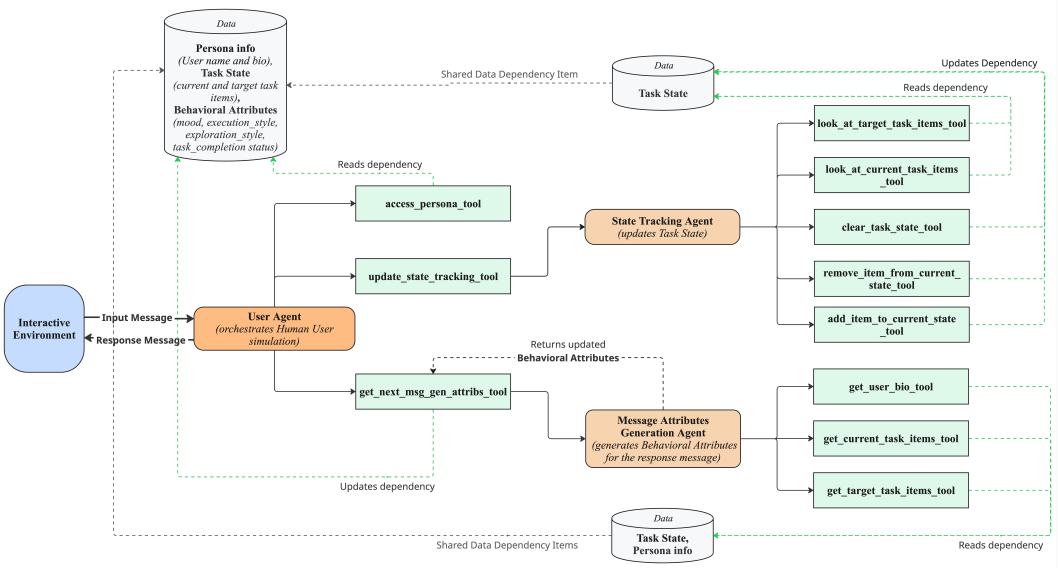

Figure 2: **Multi-agent architecture for human user simulation** showing the three-agent framework: (1) **User Agent** serves as the primary orchestrator that generates simulated user responses by receiving input messages and invoking the two sub-agents in sequence, (2) **State Tracking Agent** maintains 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.

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 $\rightarrow$ 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%

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

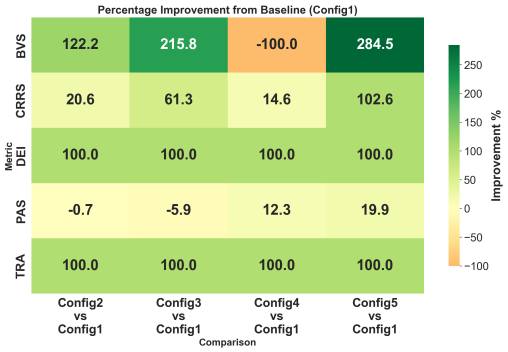

Figure 3: Performance gains from baseline

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

4.4.1 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.

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.

4.4.2 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.

Acknowledgements

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.

References

[1] Pydanticai documentation. `https://ai.pydantic.dev/`, 2024. Accessed: 2025-09-01.

[2] Adnan Ahmad, Stefan Hillmann, and Sebastian Möller. Simulating User Diversity in Task-Oriented Dialogue Systems using Large Language Models. In *arXiv preprint arXiv:2502.12813*, 2025.

[3] Krisztian Balog and ChengXiang Zhai. User Simulation in the Era of Generative AI: User Modeling, Synthetic Data Generation, and System Evaluation. In *arXiv preprint arXiv:2501.04410v1*, 2025.

[4] Nolwenn Bernard and Krisztian Balog. Towards a Formal Characterization of User Simulation Objectives in Conversational Information Access. In *arXiv preprint arXiv:2406.19007v1*, 2024.

[5] Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. PERSONA: A Reproducible Testbed for Pluralistic Alignment. In *arXiv preprint arXiv:2407.17387*, 2024.

[6] Yi Cheng, Wenge Liu, Kaishuai Xu, Wenjun Hou, Yi Ouyang, Chak Tou Leong, Wenjie Li, Xian Wu, and Yefeng Zheng. AutoPal: Autonomous Adaptation to Users for Personal AI Companionship. In *arXiv preprint arXiv:2406.13960*, 2024.

[7] KuanChao Chu, Yi-Pei Chen, and Hideki Nakayama. Cohesive Conversations: Enhancing Authenticity in Multi-Agent Simulated Dialogues. In *arXiv preprint arXiv:2407.09897*, 2024.

[8] Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. Towards Multimodal Emotional Support Conversation Systems. In *arXiv preprint arXiv:2408.03650*, 2024.

[9] Yufan Dang, Chen Qian, Xueheng Luo, et al. Multi-Agent Collaboration via Evolving Orchestration. In *arXiv preprint arXiv:2505.19591*, 2025.

[10] Sam Davidson, Salvatore Romeo, Raphael Shu, James Gung, Arshit Gupta, Saab Mansour, and Yi Zhang. User Simulation with Large Language Models for Evaluating Task-Oriented Dialogue. In *arXiv preprint arXiv:2309.13233*, 2023.

[11] Rishikesh Devanathan, Varun Nathan, and Ayush Kumar. Why Synthetic Isn't Real Yet: A Diagnostic Framework for Contact Center Dialogue Generation. In *arXiv preprint arXiv:2508.18210*, 2025.

[12] Shutong Feng, Hsien chin Lin, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. Emotionally Intelligent Task-oriented Dialogue Systems: Architecture, Representation, and Optimisation. In *arXiv preprint arXiv:2507.01594*, 2025.

[13] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling Synthetic Data Creation with 1,000,000,000 Personas. In *arXiv preprint arXiv:2406.20094*, 2024.

[14] Pengbo Hu and Xiang Ying. Unified Mind Model: Reimagining Autonomous Agents in the LLM Era. In *arXiv preprint arXiv:2503.03459*, 2025.

[15] Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, et al. GPT-4o System Card. In *arXiv preprint arXiv:2410.21276*, 2024.

[16] Jinghan Jia, Abi Komma, Timothy Leffel, Xujun Peng, Ajay Nagesh, Tamer Soliman, Aram Galstyan, and Anoop Kumar. Leveraging LLMs for Dialogue Quality Measurement. In *arXiv preprint arXiv:2406.17304*, 2024.

[17] Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. OrchestraLLM: Efficient Orchestration of Language Models for Dialogue State Tracking. In *arXiv preprint arXiv:2311.09758*, 2024.

[18] Elad Levi and Ilan Kadar. IntellAgent: A Multi-Agent Framework for Evaluating Conversational AI Systems. In *arXiv preprint arXiv:2501.11067*, 2025.

[19] Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. LLM Generated Persona is a Promise with a Catch. In *arXiv preprint arXiv:2503.16527v1*, 2025.

[20] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as Agents. In *arXiv preprint arXiv:2308.03688*, 2023.

[21] Subhankar Maity and Aniket Deroy. Generative AI and Its Impact on Personalized Intelligent Tutoring Systems. In *arXiv preprint arXiv:2410.10650*, 2024.

[22] Shuhaib Mehri, Xiaocheng Yang, Takyoung Kim, Gokhan Tur, Shikib Mehri, and Dilek Hakkani-Tür. Goal Alignment in LLM-Based User Simulators for Conversational AI. In *arXiv preprint arXiv:2507.20152*, 2025.

[23] Lingbo Mo, Shun Jiang, Akash Maharaj, Bernard Hishamunda, and Yunyao Li. HierTOD: A Task-Oriented Dialogue System Driven by Hierarchical Goals. In *arXiv preprint arXiv:2411.07152*, 2024.

[24] Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. Evaluation and Benchmarking of LLM Agents: A Survey. In *arXiv preprint arXiv:2507.21504v1*, 2025.

[25] Maria Molchanova, Anna Mikhailova, Anna Korzanova, Lidiia Ostyakova, and Alexandra Dolidze. Exploring the Potential of Large Language Models to Simulate Personality. In *Workshop on Customizable NLP (CustomNLP4U) at EMNLP 2024*, 2025.

[26] Cheng Niu, Xingguang Wang, Xuxin Cheng, Juntong Song, and Tong Zhang. Enhancing Dialogue State Tracking Models through LLM-backed User-Agents Simulation. In *arXiv preprint arXiv:2405.13037*, 2024.

[27] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. In *arXiv preprint arXiv:2304.03442*, 2023.

[28] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative Agent Simulations of 1,000 People. In *arXiv preprint arXiv:2411.10109*, 2024.

[29] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie J. Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Simulating human behavior with ai agents. Policy brief, Stanford Institute for Human-Centered Artificial Intelligence (HAI), May 2025.

[30] Vitou Phy, Yang Zhao, and Akiko Aizawa. 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] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset. In *Proceedings of AAAI 2020*, 2020.

[33] Shaina Raza, Ranjan Sapkota, Manoj Karkee, and Christos Emmanouilidis. 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] Arpita Saggar, Jonathan C. Darling, Vania Dimitrova, Duygu Sarikaya, and David C. Hogg. Score Before You Speak: Improving Persona Consistency in Dialogue Generation using Response Quality Scores. In *Proceedings of ECAI 2025*, 2025.

[35] Raphael Shu, Nilaksh Das, Michelle Yuan, Monica Sunkara, and Yi Zhang. Towards Effective GenAI Multi-Agent Collaboration: Design and Evaluation for Enterprise Applications. In *arXiv preprint arXiv:2412.05449*, 2024.

[36] Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive Architectures for Language Agents. In *arXiv preprint arXiv:2309.02427*, 2024.

[37] Weiwei Sun, Shuyu Guo, Shuo Zhang, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. Metaphorical User Simulators for Evaluating Task-oriented Dialogue Systems. In *arXiv preprint arXiv:2204.00763*, 2022.

[38] Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and Eng Siong Chng. DiaSynth: Synthetic Dialogue Generation Framework for Low Resource Dialogue Applications. In *Proceedings of NAACL 2025*, 2025.

[39] Richard Sutcliffe. A Survey of Personality, Persona, and Profile in Conversational Agents and Chatbots. In *arXiv preprint arXiv:2401.00609*, 2023.

[40] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. In *arXiv preprint arXiv:2501.06322*, 2025.

[41] Hiromi Wakaki, Yuki Mitsufuji, Yoshinori Maeda, Yukiko Nishimura, Silin Gao, Mengjie Zhao, Keiichi Yamada, and Antoine Bosselut. ComperDial: Commonsense Persona-grounded Dialogue Dataset and Benchmark. In *arXiv preprint arXiv:2406.11228*, 2024.

[42] Hongru Wang, Min Li, Zimo Zhou, Gabriel Pui Cheong Fung, and Kam-Fai Wong. KddRES: A Multi-level Knowledge-driven Dialogue Dataset for Restaurant Towards Customized Dialogue System. In *arXiv preprint arXiv:2011.08772*, 2020.

[43] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhi-Yuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A Survey on Large Language Model based Autonomous Agents. In *arXiv preprint arXiv:2308.11432*, 2025.

[44] Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. Humanoid Agents: Platform for Simulating Human-like Generative Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 167–176, Singapore, 2023. Association for Computational Linguistics.

[45] Yafei Xiang, Hanyi Yu, Yulu Gong, Shuning Huo, and Mengran Zhu. Text Understanding and Generation Using Transformer Models for Intelligent E-commerce Recommendations. In *arXiv preprint arXiv:2402.16035*, 2024.

[46] Qiuejie Xie, Qiming Feng, Tianqi Zhang, et al. Human Simulacra: Benchmarking the Personification of Large Language Models. In *Proceedings of ICLR 2025*, 2025.

[47] Lin Xu, Ningxin Peng, Daquan Zhou, See-Kiong Ng, and Jinlan Fu. Chain of Thought Explanation for Dialogue State Tracking. In *arXiv preprint arXiv:2403.04656*, 2024.

[48] Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. A Survey on Recent Advances in LLM-Based Multi-turn Dialogue Systems. In *arXiv preprint arXiv:2402.18013*, 2024.

[49] Huizi Yu, Jiayan Zhou, Lingyao Li, et al. Simulated patient systems are intelligent when powered by large language model-based AI agents. In *arXiv preprint arXiv:2409.18924*, 2024.

[50] Wentao Zhang, Liang Zeng, Yuzhen Xiao, et al. AgentOrchestra: A Hierarchical Multi-Agent Framework for General-Purpose Task Solving. In *arXiv preprint arXiv:2506.12508*, 2025.

[51] Jiachen Zhu, Menghui Zhu, Renting Rui, Rong Shan, Congmin Zheng, Bo Chen, Yunjia Xi, Jianghao Lin, Weiwen Liu, Ruiming Tang, Yong Yu, and Weinan Zhang. Evolutionary Perspectives on the Evaluation of LLM-Based AI Agents: A Comprehensive Survey. In *arXiv preprint arXiv:2506.11102v1*, 2025.

[52] Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, Fazl Barez, Rahul Gupta, Jwala Dhamala, Jacob Merizian, Mario Giulianelli, Harry Coppock, Cozmin Ududec, Jasjeet Sekhon, Jacob Steinhardt, Antony Kellerman, Sarah Schwettmann, Matei Zaharia, Ion Stoica, Percy Liang, and Daniel Kang. Establishing Best Practices for Building Rigorous Agentic Benchmarks. In *arXiv preprint arXiv:2507.02825v2*, 2025.

[53] Mingchen Zhuge, Changsheng Zhao, Dylan R. Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-Judge: Evaluate Agents with Agents. In *arXiv preprint arXiv:2410.10934*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our main contributions: (1) a novel multi-agent framework for human user simulation with specialized agents, (2) systematic evaluation methodology with ablation studies, and (3) comprehensive test dataset in restaurant ordering domain. The experimental results provided in the paper directly support these claims, and limitations are explicitly discussed.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 explicitly discusses the limitations of the work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not include theoretical results that require formal proofs. The work is primarily empirical, focusing on a multi-agent system architecture and its experimental validation through ablation studies.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed experimental setup including 5 ablation configurations, implementation details using Pydantic AI with GPT-4o, evaluation metrics with mathematical formulations, and comprehensive results with statistical significance testing.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The data and the code will be open sourced and as supplementary material, the codebase and data for this complete work will be provided.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all relevant experimental details: 60 test cases per configuration, GPT-4o model usage, Pydantic AI implementation framework, conversation management with turn limits and repetition detection, and comprehensive logging for evaluation. No training is involved as this uses pre-trained models. The order test case data generation prompts can also be found in the codebase provided as the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Table 3 provides statistical significance analysis with p-values for all metrics comparing the full system vs baseline. All improvements are statistically significant ($p < 0.01$), with specific significance levels marked (** $p < 0.01$, *** $p < 0.001$).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resource specs are provided in the experimental setup section of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research involves only synthetic data generation and agent simulation without human subjects. Ethical considerations are explicitly addressed in Section 5, including transparency requirements, bias prevention, and responsible use guidelines for user simulation systems.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 5 discusses positive impacts (democratizing conversational AI testing, systematic quality assurance) and negative impacts (potential for impersonation without consent, manipulation risks, bias perpetuation). Future applications across healthcare, education, and customer service are also outlined.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release pretrained models, scraped datasets, or other high-risk assets. The framework uses existing models (GPT-4o) and synthetic datasets (restaurant personas and menu items) that pose minimal misuse risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites GPT-4o with appropriate references and acknowledges the use of Pydantic AI framework. All external datasets, models, and libraries referenced in the methodology are properly attributed through citations.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces new synthetic datasets (20 personas, restaurant menu, 60 test cases) and evaluation metrics (PAS, BVS, TRA, DEI, CRRS) which are thoroughly documented in the methodology section with mathematical formulations and clear definitions.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This research does not involve crowdsourcing or human subjects. All experiments are conducted using synthetic data and AI agent simulations without human participation.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This research does not involve human subjects or crowdsourcing, so IRB approval is not required. All experiments are conducted using AI agents and synthetic data.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs (specifically GPT-4o) are a core component of the methodology, serving as the underlying models for all three agents in the multi-agent framework. The paper clearly describes their usage in the implementation section and throughout the experimental setup.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.