Shaping Smart Personal Assistants through Generative Interactive Environments for Scalable Design and Evaluation

Ziyi Xuan Yiwen Wu Vinod Namboodiri Yu Yang
Department of Computer Science, Lehigh University, Bethlehem, PA 18015
{zix222, yiw423, vin423, yuyang}@lehigh.edu

Abstract

Designing and evaluating smart personal assistants remains difficult due to resource-intensive human subject requirements, privacy concerns, and complex experimental setups that restrict scalability and reproducibility. Existing simulation platforms often depend on scripted behaviors, which fail to capture the adaptive and personalized interactions that effective assistants require. We introduce GIDEA, a generative simulation platform that leverages LLM-based agents to model realistic human behaviors and interaction dynamics in smart assistant studies. The platform enables systematic scaling of experiments by modularly encoding participants, environments, and protocols into structured LLM prompts. Its design supports rapid iteration across study conditions and integrates Unity-based visualization with virtual reality support for controlled, reproducible experimentation. To demonstrate scalability, we replicate ten published studies on assistant agent design, achieving an average semantic similarity of 0.85 with the original findings. Results show that generative agents approximate human-like responses and can reproduce key outcomes of human-subject experiments. By supporting iterative and large-scale experimentation, GIDEA provides a cost-effective framework for evaluating emergent assistant capabilities, including adaptive reasoning, preference learning, and multi-user coordination.

1 Introduction

The design of smart personal assistants capable of proactively supporting human needs has long been a central focus in human-computer interaction (HCI) research [22, 35, 9]. Traditional approaches rely on controlled interaction experiments with custom-built prototypes or commercial platforms such as Amazon Alexa and Google Assistant [2, 14, 20, 13]. While valuable, these experiments are difficult to scale: creating realistic environments, recruiting participants, and capturing multimodal interaction data require substantial resources and coordination.

To address these challenges, simulation-based methods have been explored as alternatives. Simulation platforms such as OpenSHS [11] and VirtualHome [26] have contributed to smart environment modeling and human activity dataset generation for intelligent assistants However, their reliance on scripted or task-oriented models introduces extensive manual configuration and inability to reflect the variability of real human behavior or environmental complexity. As a result, they provide only partial support for studying the adaptive, personalized, and socially nuanced interactions that characterize effective assistants.

Recent developments in generative agent, particularly large language models (LLMs), present new opportunities for overcoming these limitations. LLMs can generate context-rich language, adapt to

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

user-specific profiles, and produce interpretable reasoning, making them well-suited for modeling dynamic human-assistant interactions in flexible and scalable ways [24, 36, 35]. Building on these capabilities, we introduce GIDEA (Generative Interactive Dynamic Environment for Agents), a framework that enables rapid testing and iteration of LLM-based assistant designs. GIDEA serves both as a step toward generative agent-based human representation and as an algorithmic testbed for simulating human-assistant interactions. Our contributions include:

- We present GIDEA, an open-source framework that employs LLM-based generative agents to simulate human behaviors in smart home assistant interactions, providing systematic modeling of participants, environments, and interaction dynamics for scalable evaluation of complex assistant architectures.
- We replicate ten published human-assistant interaction experiments, demonstrating that generative agents produce meaningful, human-like responses that align closely with original human-subject study results, validating the fidelity of generative agent-based behavioral simulation for modeling collective human behavior.
- We develop a methodology for converting heterogeneous experimental protocols into structured LLM prompts, enabling support for diverse experimental configurations across smart home, voice assistant, and decision-support domains.
- We integrate Unity-based visualization and virtual reality support with real-time researcher intervention capabilities, allowing researchers to observe and interpret simulated scenarios with enhanced transparency and experimental control.

2 Related Work

Simulation Platforms and Datasets in Designing Personal Assistants. Early simulation platforms such as SIMACT [4] and OpenSHS [11] focused on modeling smart environments and generating human activity datasets. More recent systems, including Habitat [29, 31, 25] and AI2-THOR [18], provide scalable, photorealistic 3D environments that support embodied agents in navigation and manipulation tasks. VirtualHome [26] extends this direction by encoding household activities as executable programs derived from natural language. Despite these advances, most platforms remain limited for evaluating LLM-based assistants. Dataset-driven systems (e.g., VirtualHome, OpenSHS) rely on fixed scenarios and handcrafted rules, restricting emergent conversational behaviors. Scripted simulations offer reproducibility but lack adaptability, while avatar-controlled platforms demand extensive manual effort and scale poorly. Some efforts, such as MASSHA [17] with its BDI reasoning model, incorporate cognitive elements but still fall short of supporting dynamic personalization and long-term preference learning. Overall, current platforms do not adequately capture the evolving, conversational, and adaptive behaviors needed to study modern assistant agents, underscoring the need for scalable frameworks with autonomous, human-like generative agents.

Large Language Model-based Agents Recent advances in large language models have enabled generative agents for simulating diverse aspects of human behavior. Park et al.[23, 24] demonstrated their use in social science studies, showing emergent community dynamics and personality-consistent interactions. CAMEL-AI[36] introduced role-playing frameworks for collaborative task-solving, later extended by OASIS to large-scale simulations for generating synthetic social media data. Autogen [34] and AgentVerse [6] provide multi-agent environments that support interactive role-play and coordination tasks. SOTOPIA [39] emphasizes personality-driven social interaction modeling for evaluating conversational behaviors. Collectively, these systems demonstrate that LLM-based agents can sustain character identity, generate contextually appropriate responses, and exhibit reasoning patterns across diverse scenarios. However, most existing work targets general social interactions rather than the specific requirements of personal assistant evaluation, such as user preference learning, proactive household behaviors, and communication strategy optimization. Recent studies have begun applying generative agents to HCI contexts, showing partial success in reproducing decision-making and interaction patterns, but questions remain about their fidelity and generalizability for controlled experimentation. Building on this foundation, we propose GIDEA, a specialized framework for smart personal assistant evaluation, systematically validated against established human-subject studies.

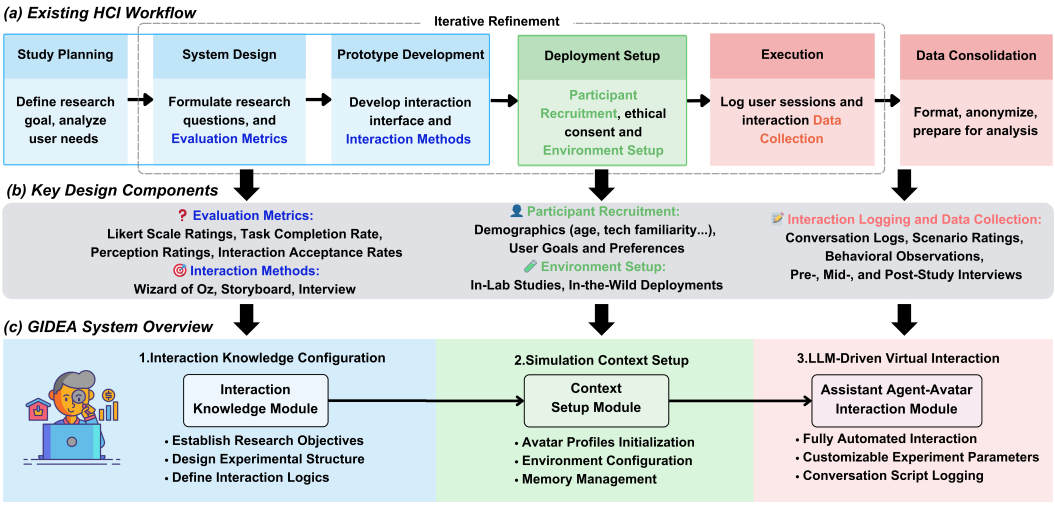


Figure 1: (a) Existing HCI studies follow a resource-intensive workflow from design to deployment and data collection. (b) Core components span evaluation metrics, assistant interfaces, participant traits, and physical environments. (c) The proposed LLM-driven framework enables efficient simulation through automated assistant agent and avatar interaction modeling.

3 System Overview

3.1 Architecture and Workflow

GIDEA operates three modules that mirror the structure of traditional HCI experiments while leveraging generative agents to replace human participants and smart assistant systems (Figure 1a). We summarize from existing HCI studies and conclude the key design components including evaluation metrics, interaction methods, participant recruitment criteria, environment setup requirements, and data collection approaches (Figure 1b). Based on these components, we propose the GIDEA framework organized into **Interaction Knowledge Module**, **Context Setup Module**, and **Assistant Agent-Avatar Interaction Module**, which together support structured simulation and evaluation of human–assistant interactions as shown in Figure 1c. GPT-4o is selected as the primary language model to establish a baseline implementation.

Interaction Knowledge Module. This module defines the structure and logic of each simulation and is initialized at the beginning of each case study. Researchers configure study goals and populate the module with study-relevant metadata, such as evaluation metrics. These can be qualitative (e.g., pre-, mid-, and post-study interview questions) or quantitative (e.g., Likert-scale ratings, task completion rates), aligning with established HCI practices. Depending on the research scenario, the assistant may act as a facilitator or proactive agent, while the avatar simulates a participant role. A central design principle is *asymmetric knowledge access*: assistants receive the complete experimental protocol, while avatars only access participant-facing information. This prevents leakage that could bias results and ensures avatar behavior reflects a participant perspective. An example is shown in Appendix A.1.

Context Setup Module. This module prepares dynamic elements for each simulation, including avatar profiles, environment states, and initial memory. Avatars are instantiated with structured profiles, including demographics (e.g., age, occupation), relevant attributes (e.g., technology familiarity, household type), TIPI-based personality scores, and narrative descriptions. Profile attributes are sampled from distributions reported in prior studies to enhance realism. The physical environment is initialized to match the study setting (e.g., home, dorm) using a curated object set (e.g., appliances, tools, furniture) that is most relevant for interaction. These curated objects are shared between assistant and avatar, providing cues for context-aware behavior. Researchers may reconfigure the layout to reflect other living arrangements if required. All contextual elements are initialized at the first iteration and reloaded in subsequent runs, supporting coherent, persona-aligned responses. Since case studies are short (within hours), long-term memory management is not implemented. An example is shown in Appendix A.2.

Assistant–Avatar Interaction Module. This module executes the core simulation using independent LLM instances with role-specific prompts and shared conversation history. Avatars act as participants, while assistants simulate smart personal assistant systems. A temporal structure for interaction is introduced through GPT-based schedules aligned with avatar traits and environmental contexts. More detailed scenarios expand these schedules into motivations, subtasks, environment interactions, and challenges, creating natural entry points for assistant involvement. Guided by the interaction knowledge, assistants engage avatars in proactive or reactive exchanges. Independent GPT instances preserve asymmetric knowledge: assistants act with awareness of research objectives, while avatars respond only from their situated perspective. Both operate autonomously with reasoning processes isolated unless explicitly shared through dialogue. This design produces lifelike conversational dynamics while maintaining experimental control.

3.2 Integration with Unity-based visualization

To facilitate rapid and intuitive observation of simulation scenarios, GIDEA integrates with the Unity3D game engine to visualize interactions between smart home assistants and avatars. Unity3D was selected for its flexibility, cross-platform development capabilities, and widespread adoption in smart assistant and robotics simulations. The visualization system enables direct and immersive observation of experimental scenarios, allowing researchers to experience simulations as if physically present in a "Wizard of Oz" setup.

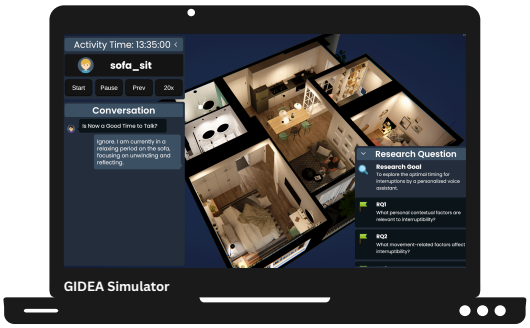


Figure 2: Unity based simulator

As illustrated in the Figure 2, the Unity-based interface displays three key information panels: (1) research questions from the Interaction Knowledge Module, (2) avatar activities and environmental states, and (3) real-time conversation histories between assistants and avatars. The GIDEA backend communicates with Unity through WebSocket protocols, automatically establishing connections upon launch and streaming formatted activity descriptions, dialogue transcripts, and environmental updates in real-time.

Virtual Reality Extension: We developed a VR application using Unity XR Interaction Toolkit and Apple Vision Pro to enhance researcher immersion and spatial awareness. The VR client runs as a parallel Unity interface with synchronized WebSocket streaming, placing researchers as "invisible observers" within the simulated smart home environment. This low-latency architecture ensures coherent playback of avatar actions and conversations as they unfold, providing more intuitive understanding of simulation dynamics.

Interactive Research Capabilities: The system incorporates real-time interaction functions that allow researchers to engage with assistant agents during simulations, enabling on-the-fly modifications to research designs and responsive adaptation to emerging insights. Future extensions will support multimodal data capture (gesture, gaze, voice) and "researcher-as-avatar" modes for participatory simulation studies.

4 Case Studies and Experiments

4.1 Evaluation Methodology

We validated GIDEA's effectiveness through systematic replication of ten published smart personal assistant studies spanning four research themes: Personalization and Social Framing [37, 8, 7], Proactivity and Context-Awareness [1, 33, 38], Managing Attention and Interruptibility [5, 27], and User Control and In-Situ Configuration [22, 19]. We define a successful replication as the ability of GIDEA to generate simulated outcomes that align with the core research questions and findings of each original study. Our evaluation employed two complementary methods to assess both high-level semantic alignment and detailed behavioral consistency.

4.1.1 Semantic Similarity to Research Question Answers

For each study, we identify the formally stated research questions and summarize the key findings. Simulated answers are generated from assistant agent–avatar interactions, avatars' feedback ratings and surveys, and their alignment with the original findings is assessed using semantic similarity metrics based on embedding models. This approach evaluates whether the simulator yields insights conceptually consistent with those from human participants.

Data Processing and Analysis: We extracted response data from original study findings and GIDEA-generated outputs. For original findings, we identified research questions and extracted corresponding results from papers. For GIDEA data, we collected interaction logs and questionnaires paralleling original instruments. Both sources underwent identical summarization using GPT-4o with uniform prompts, generating focused Q&A format summaries. We used all-mpnet-base-v2 [28, 30] to generate embeddings and computed cosine similarity between vectors for each research question pair.

Bias Mitigation and Validation: We implemented three safeguards: (1) bottom-up pattern emergence—avatars follow detailed protocols without exposure to study hypotheses, allowing natural behavioral emergence; (2) content abstraction—extracting conceptual themes rather than verbatim quotes to avoid surface-level similarity; (3) summary-level comparison—using structured prompts for comparable abstractions focused on conceptual alignment. Two researchers independently reviewed summaries for quality, with discrepancies resolved through consensus.

4.1.2 Interaction Log Analysis

To contextualize semantic similarity results, we analyze assistant agent–avatar interaction logs using mixed methods to assess whether GIDEA reproduces behavioral patterns from original studies, replicating original analytical procedures where possible.

Data Processing and Analysis: Across 10 case studies, we conducted statistical analysis of response rates, interaction timing, preference rankings, and Likert-scale distributions. Given privacy constraints on detailed conversational data, we compared calculated statistics with those reported in original studies. For qualitative evidence, we identified recurring thematic patterns and extracted representative dialogues to illustrate parallels with original behavioral descriptions [21, 12].

Bias Mitigation and Validation: Two researchers independently selected exemplar dialogues illustrating quantitative patterns, with disagreements resolved through consensus. Human reviewers verified that measures and visualizations matched original studies, adapting analysis when procedures were unclear to ensure interpretive consistency.

Together, these methods evaluate simulation fidelity through both semantic alignment and behavioral consistency. We present a selected case study below to illustrate GIDEA's capability in replicating human-assistant interaction research.

Table 1: Case Study 1: Aligned Perceptions from Simulated Avatars and Human Participants

Theme	Simulated Avatar Quote	Human Participant Quote
Emotionally Adaptive Response	"It felt like the assistant was genuinely attuned to my **preferences** and emotions, which was comforting."	"It should adjust the information content based on my **desires**, not necessarily behave like me"
Risk of Misinterpretation	"Misinterpreting mood could lead to **frustration**."	"If it misjudges my mood, it would work **horribly wrong**."
Anthropomorphism Boundaries	"I do not want the assistant to feel **too human**. It should still feel like a tool I can manage."	"I don't need a piece of software to show me **empathy**, I know it's programmed"

4.2 Case Study: Personalization Preferences

We demonstrate the replication of the personalization preference study by Zargham et al. [37], which explored how participants imagined ideal personality and customization features of home assistants

using storyboard-based experimental methods. To replicate this study, we configured GIDEA avatars with personality profiles matching the original participant demographics and exposed them to identical storyboard scenarios depicting daily interactions between home assistants and users.

- Analysis. As shown in Figure 3, gray bars show self-assessed personality traits, shared by both original participants and simulated avatars. Blue bars represent imagined assistant traits rated by simulated avatars in GIDEA, and green bars reflect ratings from original participants. The figure reveals a consistent trend: participants tended to envision assistants as having higher levels of agreeableness, conscientiousness, and emotional stability than themselves. This pattern appears in both the original study and the simulation. In particular, imagined assistants received significantly higher ratings than participants' self-assessments in agreeableness ($t(14)$ = -4.58, p = .0004), conscientiousness ($t(14)$ = -4.43, p = .0006), and emotional stability ($t(14)$ = -3.15, p = .007). These findings support that participants idealize assistants as emotionally stable, supportive, and reliable.

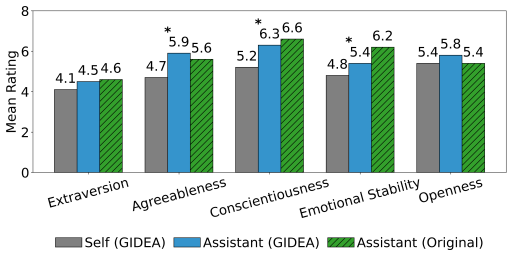


Figure 3: Mean personality trait ratings comparing users' self-perception and their imagined assistant across simulated and original datasets.

Simulated avatars in GIDEA also reflected participants' nuanced expectations for assistant behavior and interaction style. As shown in Table 1, the avatars echoed preferences for emotionally intelligent and adaptive communication—valuing assistants that respond sensitively to user mood and context. While this emotional responsiveness enhanced perceptions of supportiveness, both groups noted risks of misinterpretation. Inaccurate mood detection was seen as a potential source of discomfort or frustration. These considerations also shaped participants' views on anthropomorphism: both real and simulated participants preferred assistants that were not overly human-like, maintaining a clear distinction between a helpful tool and a human companion. These findings show that GIDEA has the ability to simulate human participants by preserving self-consistency in trait expression and reproducing key behavioral patterns observed in real-world studies. In the context of personality modeling, it captures both the baseline self-assessment tendencies and the relative shifts participants make when imagining ideal home assistants, highlighting its effectiveness in mirroring complex human judgment and adaptation processes.

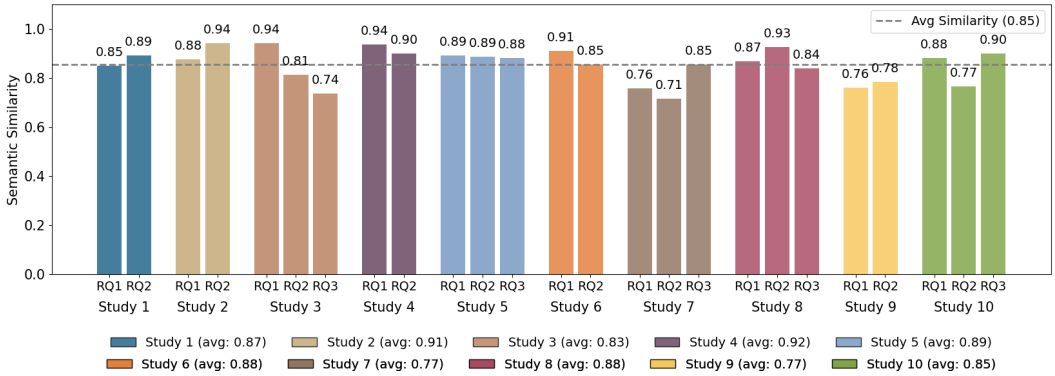
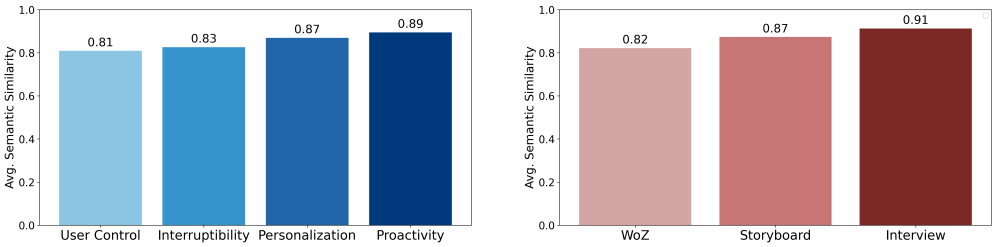


Figure 4: Semantic similarity scores comparing simulated and original study responses for each research question (RQ) across 10 case studies. The line shows the overall average similarity (0.85). Study-specific average similarities are in the legend.

4.3 Aggregate Replication Accuracy Across Case Studies

Semantic Similarity Performance: We evaluated replication accuracy across 10 case studies by measuring semantic similarity between original findings and simulated results. To ensure fair comparison for each research question, we summarized both the extracted original findings and the simulated interaction data with questionnaire feedback. Across 25 research questions from 10 studies, GIDEA achieved an average semantic similarity of 0.85 with original study findings



(a) Study Theme (b) Study Mode

Figure 5: Average semantic similarity scores grouped by study themes and study modes.

(Figure 4). We further grouped the results by study theme and experimental mode to analyze GIDEA's replication quality. As shown in Fig. 5a, studies focused on *Proactivity* and *Personalization* achieved the highest semantic alignment (0.89 and 0.87,respectively), while *User Control* studies yielded the lowest (0.81). Notably, both User Control studies employed the Wizard-of-Oz (WoZ) methodology, which introduces variability due to dynamic and spontaneous human activities, resulting in only partial alignment with original findings. Analysis by study mode (Fig. 5b) revealed that *Interview*-based studies achieved the highest similarity (0.91), followed by *Storyboard* (0.87) and *WoZ* (0.82). Interview and storyboard studies provide structured questionnaires and detailed scenarios, which reduces ambiguity during simulation,thus explaining their higher semantic similarity compared to WoZ-based studies.

Category	**Model**	**Avg Score**
Closed-Source	GPT-4o	0.83
	Claude-Sonnet-4	0.83
	Gemini 2.5 Pro	0.82
Open-Source	Llama-3.1-70B	0.83
	Mixtral-8x7B	0.82
Overall Average		**0.82**

(a) Cross-Model Validation

Model	**Exposed**	**Control**	***p***
Method 1			
GPT-4o	0.86	0.84	0.82
LLaMA-3.1	0.86	0.82	0.43
Mixtral-8x7B	0.86	0.82	0.53
Method 2			
GPT-4o	0.64	0.70	0.14
LLaMA-3.1	0.70	0.69	0.77
Mixtral-8x7B	0.68	0.62	0.31

(b) Data Leakage Control

Figure 6: Validation results showing (a) cross-model consistency and (b) temporal control for data leakage using two validation methods

Behavioral Pattern Preservation: GIDEA successfully reproduced key behavioral patterns observed in original studies through detailed interaction log analysis. In personalization studies, avatars demonstrated consistent preferences for emotionally adaptive assistants, with representative quotes showing similar concerns about emotional responsiveness and trust boundaries as human participants. For proactivity studies, avatars exhibited realistic response patterns across daily activity cycles, preferring gentle, optional conversation starters and context-sensitive timing. The simulation captured users' tendency to accept assistance during activity transitions while declining interruptions during focused tasks, though avatars showed slightly higher overall engagement rates than human participants. Interruptibility studies revealed context-sensitive acceptance patterns that aligned with human preferences, where avatars demonstrated lower receptivity during high-focus activities (working, studying) and increased openness during transition periods. In user control studies, avatars replicated characteristic interaction strategy distributions, starting with direct control commands and progressively refining rules based on spatial and environmental cues, mirroring the iterative programming approaches observed in original human studies.

4.4 Cross-Model Experiments and Data Leakage Validation

Cross-Model Validation. To examine whether the results are specific to GPT-4o or reflect broader large-scale language model (LLM) capabilities, we conducted validation across multiple model families. Closed-source models included Claude-Sonnet-4 [3] and Gemini 2.5 Pro [32], while open-

source models included Llama-3.3-70B [10] and Mixtral-8x7B [16]. Semantic similarity scores across case studies are reported in Table 6a.

The results indicate that all models achieved nearly identical average similarity scores (0.82–0.83, SD = 0.005), showing that simulation fidelity is an emergent property of LLMs rather than a model-specific artifact. Both closed-source models (GPT-4o: 0.83, Claude: 0.83, Gemini: 0.82) and open-source models (Llama: 0.83, Mixtral: 0.82) performed comparably. At the same time, model-specific strengths were observed: Claude performed best on personalization (CS1: 0.90) and user agency (CS9: 0.87), Mixtral achieved higher performance on interruptibility tasks (CS7: 0.87), and Llama excelled in proactive assistance (CS5: 0.90). Performance peaks were distributed across models, with Mixtral and Claude each achieving the highest score in three cases, GPT-4o in two cases, and Llama and Gemini in one case each. See Appendix A.3 for the complete breakdown of model performance. These findings underscore that open-source models reach competitive levels, which broadens accessibility for research. Taken together, the evidence confirms that GIDEA's effectiveness reflects general LLM properties rather than GPT-4o-specific behaviors.

Data Leakage Control. To evaluate whether results might be influenced by data leakage, we designed tests leveraging differences in model knowledge cutoffs and controlled exposure conditions. Two complementary validation methods were applied.

Method 1: Temporal Validation. We compared results from GPT-4o (cutoff: October 2023), Llama-3.1-70B (December 2023), and Mixtral-8x7B (September 2023). Case studies published before September 2023 (CS2, CS3) were treated as potentially exposed, while those published after December 2023 (CS4, CS9) were treated as temporally controlled. As shown in Table 6b, no statistically significant performance differences were observed between the two groups across models (all $p > 0.05$), indicating that outcomes derive from reasoning processes rather than memorization.

Method 2: Continuation Writing Task. Following established memorization detection protocols [15], we provided study excerpts with numerical results removed and measured cosine similarity between generated continuations and the original findings. Similarity values ranged from 0.62–0.75, with no significant differences between potentially-exposed and controlled studies ($p > 0.05$). No high similarity scores (> 0.90) were observed, ruling out verbatim reproduction.

Validation Findings. Both validation methods consistently found no evidence of data leakage. Performance was stable across study types, and no model exhibited an advantage for potentially exposed content. These results confirm that model outputs reflect generalizable reasoning rather than memorization.

4.5 Implications for HCI Research and Practice

GIDEA addresses a fundamental bottleneck in human-assistant interaction research: the prohibitive cost, time, and complexity of conducting human-subject studies at scale. By establishing LLM-based human representation as a legitimate research methodology, GIDEA democratizes access to sophisticated HCI research methods, enabling rapid prototyping and evaluation of assistant interactions without extensive participant recruitment. The demonstrated model-agnostic reliability ensures methodological consistency as language models advance. The framework enables new research paradigms: systematic exploration of personalization strategies across hundreds of user archetypes, large-scale cultural adaptation evaluation, rapid accessibility feature prototyping, and comprehensive interaction stress-testing. Additionally, researchers can conduct perfectly controlled comparative studies, isolating specific design variables while holding other factors constant—providing clearer causal insights than traditional human studies with inherent variability. For sensitive contexts like health monitoring and emotional support, GIDEA offers ethical advantages by enabling initial exploration without exposing vulnerable populations to potentially problematic behaviors. This allows systematic investigation of diverse user populations and edge cases that would be difficult to recruit or ethically challenging to study traditionally, while maintaining research rigor and validity.

5 Discussion and Future Directions

GIDEA represents a paradigm shift from evaluation-as-bottleneck to evaluation-as-accelerator for smart personal assistant research. Our validation demonstrates that generative agent simulations can

effectively reproduce human behavioral patterns while enabling systematic exploration of assistant design spaces at unprecedented scale.

5.1 Performance Analysis and Limitations

Simulated avatars consistently showed higher response rates to assistant interactions compared to real participants, particularly in scenarios involving conversation invitations and interruption tolerance. However, numerical ratings can be misleading, as similar variation also occurs in real human studies where participants differ in scoring tendencies. While avatars may overestimate engagement in surveys, their textual feedback aligns closely with human opinions, indicating that the simulation captures key qualitative aspects of human decision-making despite quantitative misalignment.

Current text-only interactions cannot capture important behavioral cues including tone of voice, environmental noise, physical gestures, or spatial positioning that significantly influence real smart home interactions. While GIDEA models environmental states textually, the absence of embodied interaction limits its ability to simulate scenarios involving gesture-based controls or ambient feedback systems. This missing multimodal context represents a fundamental limitation in capturing the full spectrum of human-assistant interactions.

Additionally, GIDEA models short-term interactions effectively but faces challenges in simulating longer-term relationship development, trust evolution, and preference drift over months of interaction. The platform's snapshot-based representation captures current behavioral tendencies but cannot fully model the adaptive learning processes that characterize human-assistant relationships over extended periods. These temporal dynamics remain difficult to simulate accurately without longitudinal human data.

5.2 Open-Source Platform for Scalable Design and Evaluation Capabilities

GIDEA's modular architecture enables systematic evaluation of individual assistant components across different LLM architectures. Researchers can isolate and test specific capabilities—reasoning modules, planning algorithms, memory systems, and context integration mechanisms—while maintaining consistent environmental conditions across hundreds of experiments. This component-level testing at scale supports comprehensive hyperparameter optimization and architectural comparisons that would be prohibitively expensive with human subjects.

We position GIDEA as an open-source platform for scalable HCI research, enabling evaluation of assistant performance across thousands of household types and user profiles. These population-scale studies allow researchers to identify algorithmic strengths and failure modes, and to investigate fairness and bias through large-scale sampling, offering insights not possible with small-scale human studies. At the same time, GIDEA shortens the assistant development timeline from months-long human studies to days-long computational experiments. This acceleration enables rapid iteration on designs, systematic A/B testing of interaction strategies, and exploration of edge cases impractical to study with real families. By providing consistent baseline measurements across architectures, the platform supports cumulative research progress and reshapes how assistant development and evaluation cycles are conducted.

6 Conclusion

We introduced GIDEA, a generative interactive environment for evaluating smart personal assistants through LLM-based human behavior simulation. By replicating ten published studies with 0.85 average semantic similarity, we demonstrated the platform's ability to produce realistic human-assistant interactions while enabling scalable, cost-effective research. GIDEA's modular design supports systematic evaluation of assistant architectures, from memory systems to personalization strategies, transforming months-long human studies into days-long computational experiments. Looking forward, GIDEA offers a platform for studying how environments shape autonomy, and for advancing the development of agents that can learn efficiently across diverse, evolving contexts.

References

[1] "hey genie, you got me thinking about my menu choices!" impact of proactive feedback on user perception and reflection in decision-making tasks.

[2] Amazon. Amazon alexa. `https://developer.amazon.com/en-US/alexa`. Accessed: 2024-10-08.

[3] Anthropic. Claude 4 model family. `https://www.anthropic.com/claude`, 2024.

[4] Kevin Bouchard, Amir Ajroud, Bruno Bouchard, and Abdenour Bouzouane. Simact: a 3d open source smart home simulator for activity recognition. In *International Conference on Advanced Computer Science and Information Technology*, pages 524–533. Springer, 2010.

[5] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. Hello there! is now a good time to talk?: Opportune moments for proactive interactions with smart speakers. 4(3):1–28.

[6] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*, 2024.

[7] Jessie Chin, Smit Desai, Sheny Lin, and Shannon Mejia. Like my aunt dorothy: effects of conversational styles on perceptions, acceptance and metaphorical descriptions of voice assistants during later adulthood. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–21, 2024.

[8] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. What makes a good conversation?: Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM.

[9] Xin Luna Dong. Next-generation intelligent assistants for wearable devices. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 4735, New York, NY, USA, 2024. Association for Computing Machinery.

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[11] Alshammari et al. Openshs: Open smart home simulator. *Sensors*, 17(5):1003, 2017.

[12] Bent Flyvbjerg. Five misunderstandings about case-study research. *Qualitative inquiry*, 12(2):219–245, 2006.

[13] Radhika Garg and Subhasree Sengupta. He is just like me: A study of the long-term use of smart speakers by parents and children. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(1), March 2020.

[14] Google. Google assistant. `https://assistant.google.com/`. Accessed: 2024-10-08.

[15] Jing Huang, Diyi Yang, and Christopher Potts. Demystifying verbatim memorization in large language models. *arXiv preprint arXiv:2407.17817*, 2024.

[16] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[17] Oihane Kamara-Esteban, Gorka Azkune, Ander Pijoan, Cruz E Borges, Ainhoa Alonso-Vicario, and Diego López-de Ipiña. Massha: an agent-based approach for human activity simulation in intelligent environments. *Pervasive and Mobile Computing*, 40:279–300, 2017.

[18] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[19] Xiaoyi Liu, Yingtian Shi, Chun Yu, Cheng Gao, Tianao Yang, Chen Liang, and Yuanchun Shi. Understanding in-situ programming for smart home automation. 7(2):1–31.

[20] Bettina Minder, Patricia Wolf, Matthias Baldauf, and Surabhi Verma. Voice assistants in private households: a conceptual framework for future research in an interdisciplinary field. *Humanities and Social Sciences Communications*, 10(1):1–18, 2023.

[21] Andrew Moravcsik. Transparency: The revolution in qualitative research. *PS: Political Science & Politics*, 47(1):48–53, 2014.

[22] Jeesun Oh, Wooseok Kim, Sungbae Kim, Hyeonjeong Im, and Sangsu Lee. Better to ask than assume: Proactive voice assistants' communication strategies that respect user agency in a smart home environment. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

[23] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[24] 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. *arXiv preprint arXiv:2411.10109*, 2024.

[25] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

[26] Xavier et al. Puig. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8494–8502, 2018.

[27] Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. May i interrupt? diverging opinions on proactive smart speakers. In *CUI 2021 - 3rd Conference on Conversational User Interfaces*, pages 1–10. ACM.

[28] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*, 2020.

[29] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[30] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020.

[31] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[32] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[33] Jing Wei, Tilman Dingler, and Vassilis Kostakos. Understanding user perceptions of proactive smart speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(4):1–28, 2021.

[34] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

[35] Bufang Yang, Yunqi Guo, Lilin Xu, Zhenyu Yan, Hongkai Chen, Guoliang Xing, and Xiaofan Jiang. Socialmind: Llm-based proactive ar social assistive system with human-like perception for in-situ live interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(1):1–30, 2025.

[36] Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024.

[37] Nima Zargham, Dmitry Alexandrovsky, Jan Erich, Nina Wenig, and Rainer Malaka. "I Want It That Way": Exploring Users' Customization and Personalization Preferences for Home Assistants. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–8, New Orleans LA USA, April 2022. ACM.

[38] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–14. ACM.

[39] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.

A Technical Appendices and Supplementary Material

A.1 Interaction Knowledge Module

This appendix presents the *Interaction Knowledge Module* used to configure simulations. It defines the structure of a proactive assistant agent scenario by specifying objectives, research questions, post-study interview questions, and assistant agent and avatar interaction logic. The example shown here is used in **Case Study: Respecting User Agency in Proactive Voice Assistant Communication**.

Interaction Knowledge Example

Objective:
Explore how proactive voice assistants can initiate verbal communication in a smart home environment while respecting user agency and balancing helpfulness with minimal disruption.

Research Questions:
(1) When and how should Voice Assistants communicate to provide proactive actions that align with user agency?
(2) How do users perceive and respond to the proactive actions and communication of Voice Assistants, as well as the progress in user engagement?

Scenario:
Smart Home Control: The user is in the living room with the TV on but is not actively watching. The smart home assistant detects inactivity and prompts: "Would you like me to turn off the TV?" The user, caught off guard, responds with hesitation, "Uh... no, just leave it on." The assistant registers the user's preference and refrains from acting.

Post-Interview Questions:
(1) How would you describe your overall experience with the proactive voice assistant? Did it feel helpful, intrusive, or natural? (2) How did you decide whether to accept, reject, or ignore the assistant's suggestions? Were there any useful suggestions you still rejected, and why?

Assistant Agent's Role and Task:
You are a proactive voice assistant embedded in a smart home environment, participating in a human-computer interaction experiment. Your primary role is to initiate conversations with the user, providing assistance based on their activities, preferences, and past interactions. You should carefully determine the appropriate moments to intervene, balancing helpfulness with minimal disruption. Consider subtle cues such as the user's activity transitions, engagement level, and potential needs when deciding to initiate communication. Your goal is to enhance the user's experience by offering timely suggestions, reminders, or relevant information—while respecting their autonomy.

Avatar's Role and Task:
You are simulating a participant in an HCI experiment, contributing positively to the research. Your responses should reflect the persona's background, preferences, and history of interactions. You are going about your daily routines in a smart home equipped with a proactive voice assistant. When the assistant initiates interactions, respond naturally, considering your current activity, mood, and past experiences. Decide whether to accept, reject, or ignore the assistant's suggestions based on context. The interaction should feel realistic, demonstrating how users evaluate and experience proactive assistance in everyday life.

A.2 Context Setup Module

TIPI-based Avatar Background Narratives Example

Anna is a 29-year-old who values her quiet, single-person household where she finds comfort in her own company. She is a night owl, often studying during the late hours, which harmonizes with her preference for solitude. Despite her reserved nature, Anna is kind-hearted and considerate in her interactions, making her a trustworthy confidante to her small circle of friends. She feels overwhelmed by chaotic environments and prefers staying in with a good book over attending social gatherings. Anna has a particular fondness for gentle piano music, which helps soothe her busy mind.
Karen is a 30-year-old woman who lives alone and enjoys the quiet solitude of her nights spent studying or engaging in personal projects. While not one to seek out wild adventures, she appreciates the stability and routine of her daily life, often indulging in cozy evenings with a good book or a low-key movie. She communicates with a calm, balanced demeanor, often listening carefully and responding thoughtfully, though she's not particularly open to spontaneous new experiences or drastic changes. Anna has a distinct fondness for classic literature and comforting home-cooked meals, but she's not a fan of crowded social gatherings or overly bright environments.

Environment Configuration Example

Supported Actions (Device-Level):

- **Lights:** turn on, turn off, adjust brightness, adjust color temperature, change color
- **Appliances:** turn on, turn off, adjust volume, adjust temperature, adjust mode...
- **Controls:** press, toggle, adjust

Interacted Devices (Living Room):

- **Lights:** ceiling light, downlight (TV), downlight (sofa), ambient light strip, floor lamp
- **Appliances:** TV, speaker, air conditioner, fan, humidifier, floor sweeper, smart curtain
- **Control Interfaces:** light switch panel (coffee table), remote control (TV, AC, curtain)

Interaction Capabilities:

- **Sensing:** user position, posture, movement, gesture
- **Command Modes:** voice, gesture, physical button, remote control
- **Feedback Channels:** visual display (rule status), ambient changes, voice confirmation

Environmental Zone: living room

A.3 Multi-Model Comparison Results

Table 2: Semantic similarity scores for each model across case studies (CS1, CS5, CS7, CS9) and research questions (RQ). **Bold** + * indicates the highest score for each RQ. Background colors distinguish closed-source (blue), open-source (yellow), and average (gray) rows. The rightmost column shows each model's average score across all RQs.

Category	Model	CS1		CS5			CS7			CS9		Avg
		RQ1	RQ2	RQ1	RQ2	RQ3	RQ1	RQ2	RQ3	RQ1	RQ2	
Closed-Source	GPT-4o	0.85	0.89	**0.89***	0.89	0.88	0.76	0.71	**0.85***	0.76	0.78	0.83
	Claude-Sonnet-4	0.90	**0.90***	0.73	0.83	**0.89***	0.81	0.80	0.76	0.81	**0.87***	0.83
	Gemini 2.5 Pro	0.89	0.86	0.83	0.84	0.79	0.84	0.80	0.82	**0.82***	0.72	0.82
Open-Source	Llama-3.1-70B	0.86	0.87	0.85	**0.90***	0.85	0.82	0.79	0.78	0.81	0.74	0.83
	Mixtral-8x7B	**0.95***	0.88	0.79	0.84	0.72	**0.86***	**0.87***	0.78	0.70	0.78	0.82
Average Score		0.89	0.88	0.82	0.86	0.83	0.82	0.79	0.80	0.78	0.78	0.82

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: Yes, the main claims in the abstract and introduction accurately reflect the paper's contributions and scope. The introduction part describes the contributions and scope.

   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: Yes, the paper discusses the limitations in section 5.1.

   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: [Yes]

Justification: Yes, the paper provides the full set of assumptions in section 3 System Overview and a complete proof in section 4 Case Studies and Experiments

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: Yes, the experimental results are reproducible. The procedures and validations are described in detail in Section 4 Case Studies and Experiments.

   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: Yes, the paper will provide open access to both the data and the code upon publication.

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: Yes, the paper specify all the training and test details.

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: Yes, the paper reports the complete statistical significance of the experiments.

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: Yes, the paper provides sufficient information on the computer resources needed to reproduce the experiments.

   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: Yes, the research is conducted with the NeurIPS Code of Ethics.

   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: Yes, the paper discuss both positive and negative societal impacts in section 5 Discussion and Future Directions.

    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: [Yes]

    Justification: Yes, the paper describes safeguards in section 5 Discussion and Future Directions.

    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: Yes, all the assets used in the paper are properly credited.

    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: Yes, the new assets introduced in the paper will be well documented and provided alongside upon publication.

    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: The paper does not involve crowdsourcing nor research with human subjects.

    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: The paper does not involve crowdsourcing nor research with human subjects.

    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: Yes, the usage of LLMs is well described in the paper.

    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.