Design and Evaluation of Generative Agent-based Platform for Human-Assistant Interaction Research: A Tale of 10 User Studies

ZIYI XUAN, Lehigh University, USA
YIWEN WU, Lehigh University, USA
XUHAI XU, Columbia University, USA
VINOD NAMBOODIRI, Lehigh University, USA
MOOI CHOO CHUAH, Lehigh University, USA
YU YANG, Lehigh University, USA

Designing and evaluating personalized and proactive assistant agents remains challenging due to the time, cost, and ethical concerns associated with human-in-the-loop experimentation. Existing Human-Computer Interaction (HCI) methods often require extensive physical setup and human participation, which introduces privacy concerns and limits scalability. Simulated environments offer a partial solution but are typically constrained by rule-based scenarios and still depend heavily on human input to guide interactions and interpret results. Recent advances in large language models (LLMs) have introduced the possibility of generative agents that can simulate realistic human behavior, reasoning, and social dynamics. However, their effectiveness in modeling human-assistant interactions remains largely unexplored. To address this gap, we present a generative agent-based simulation platform designed to simulate human-assistant interactions. We identify ten prior studies on assistant agents that span different aspects of interaction design and replicate these studies using our simulation platform. Our results show that fully simulated experiments using generative agents can approximate key aspects of human-assistant interactions. Based on these simulations, we are able to replicate the core conclusions of the original studies. Our work provides a scalable and cost-effective approach for studying assistant agent design without requiring live human subjects. We will open source both the platform and collected results from the experiments on our website [1].

Additional Key Words and Phrases: Conversational Agent, Intelligent Personal Assistant, Intelligent User Simulation, Simulation-based Experimentation, Large Language Models, Human-Computer Interaction

1 Introduction

Designing intelligent assistants that can proactively support human needs has become a long-standing goal in human–computer interaction (HCI) research [13]. With user consent, such systems are expected to infer user states, predict needs, and deliver timely, context-aware support. To better design and improve these systems, researchers have studied human-assistant interactions through controlled experiments, using customer-built prototypes or through platforms such as Amazon Alexa and Google Assistant [2, 5, 18, 33]. However, conducting experiments in physical environments is often labor-intensive, time-consuming, and logistically complex. Practical challenges—such as setting up experimental environments, managing participants, and collecting multi-source interaction data—limit the scale, flexibility, and personalization of study designs, particularly for longitudinal deployments. Moreover, privacy and ethical considerations necessitate additional safeguards to protect participants, which in turn limit the types of interactions and scenarios that can be explored in live deployments.

To address these barriers, researchers have explored simulation-based approaches that replicate user behaviors and environmental dynamics in virtual spaces. Platforms such as OpenSHS [15] and VirtualHome [39] offer structured frameworks to model smart environments and Activities of Daily Living (ADL), which enable algorithm testing and scenario modeling in controlled, repeatable settings [46]. However, most existing simulation platforms

[1]Website for GIDEA experiment results & demo video: https://dash-gidea.github.io/

Authors' Contact Information: Ziyi Xuan, Lehigh University, USA, zix222@lehigh.edu; Yiwen Wu, Lehigh University, USA, yiw423@lehigh.edu; Xuhai Xu, Columbia University, USA, xx2489@cumc.columbia.edu; Vinod Namboodiri, Lehigh University, USA, vin423@lehigh.edu; Mooi Choo Chuah, Lehigh University, USA, mcc7@lehigh.edu; Yu Yang, Lehigh University, USA, yuyang@lehigh.edu.

are built on either manual operation or task-oriented models, which require extensive scripting and lack the flexibility to capture the variability of human behavior and physical environments. As a result, they fall short for studying human–assistant interactions, which demand personalized, adaptive, and socially complex behaviors.

Recent advances in generative AI, particularly large language models (LLMs), offer a new opportunity to enable more resource-efficient simulation approaches. LLM-based generative agents—computational entities designed to emulate believable human behaviors [36]—can generate contextually rich language, reason over multiple turns of interaction, adapt to persona attributes, and express diverse communication styles [36, 48]. Building on these capabilities, we explore how generative agents can be leveraged to simulate human–assistant interactions. Specifically, we aim to understand how generative agent-simulated humans can contribute to human–assistant interaction research, and how the generated results compare to findings from existing human subject studies. Realizing this vision introduces several practical challenges, including data limitations, variation in research setups, and restricted access to detailed behavioral records.

Challenge 1: Public datasets that link human activities to internal cognitive or contextual states are limited. Existing data sources [3, 12, 15] are often small in scale and lack semantic annotations or user-centric labels, which makes it difficult to be used to evaluate the realism of simulated behaviors directly. To overcome this limitation, we identified ten well-documented human-assistant interaction studies and aligned our simulations with their experimental protocols. These studies serve as reference points for evaluating simulation quality via behavioral metrics (e.g., strategy usage, acceptance rates) and the semantic similarity of generated responses to original research findings.

Challenge 2: Designing a platform that can accommodate the wide variation in research setups across human–assistant interaction studies (e.g., target participants, interaction techniques) is challenging. A unified format and method are needed to translate diverse study designs into consistent, operable guidelines that replace the traditional guiding role of human researchers and support automated simulation. To meet this need, we developed a modular system design that accommodates variations across all aspects of experimental protocols. The information necessary for autonomously guiding a study—such as questionnaires, task sequences, interaction strategies, and experimental rules—is encoded into structured, reusable templates. We also designed a simulation workflow that dynamically prompts the generative agent with this configured information, including an interaction memory, allowing it to manage each phase of the study and generate context-appropriate responses throughout the simulation process.

Challenge 3: Many prior HCI studies do not release full interaction logs due to privacy concerns, limiting the availability of fine-grained behavioral data for replication. To address this challenge, we carefully design an evaluation methodology that aligns simulated behaviors with reported qualitative descriptions and interaction patterns from the original studies to enable meaningful comparison in the absence of detailed logs.

In this work, we present GIDEA, the first comprehensive evaluation of generative agent–based simulation for human–assistant interactions. This paper extends our earlier demo of the GIDEA platform [53], which introduced the core system design but did not include case studies or evaluation. Grounded in the existing research workflow (details in Fig.1), we replicate ten prior user studies spanning various themes such as proactive assistance, interruptibility, adaptive personalization, and user engagement. We demonstrate that generative agents in GIDEA can reproduce key behavioral patterns observed in human-subject studies and provide a flexible way to explore additional study conditions, which potentially reduces the burden in early-stage design before involving actual human participants. We summarize our contributions as follows:

- We present GIDEA, a general-purpose simulation platform for replicating human-assistant interaction experiments using LLM-based generative agents to reenact human behaviors. The platform models participants, environments, and interactions, enabling reproducible simulation of complex user studies.

- We introduce a unified method to translate diverse experimental protocols into LLM-compatible prompts. This prompting strategy supports a wide range of experimental setups and tasks, allowing GIDEA to be applied across various domains within human-assistant interaction research, including smart homes, voice assistants, and decision-support systems.
- We replicate ten published human-assistant interaction experiments using GIDEA, where we observe that the generative agent produces meaningful, human-like responses that offer valuable insights for researchers. These responses closely align with original human-subject results, validating the fidelity and generalizability of generative agent-based behavioral simulation in controlled study settings. We will open source both the platform and collected results from the experiments on our website.

2 Related Work

2.1 Design and Evaluation of Intelligent Assistants Agent

Design Focus: The design of intelligent assistants for everyday use increasingly emphasizes personalized, context-aware, and adaptive interactions that align with user expectations and support ongoing activities. Effective interaction design requires a careful balance between automation and user control, shaped by how users perceive, engage with, and configure assistant behaviors. One important theme in prior work is the importance of personalization in tone, interaction style, and decision-making strategies [9, 11, 56]. Personalization not only improves usability but also helps build trust, especially as users often anthropomorphize assistants, attributing social roles or intentions that influence their interpretation of system behavior [9]. Another key perspective focuses on proactive engagement. Studies have examined how assistants initiate interactions based on context and timing [8, 50, 57]. While users generally value proactive support, they prefer subtle or permission-seeking cues, such as soft chimes, when assistants interrupt ongoing tasks [8, 40]. These findings point to the importance of adaptive timing strategies that maintain user agency and minimize disruption. Finally, recent research explores user-driven approaches to configuring assistant behaviors. In-situ programming methods allow users to define or refine actions during actual use, promoting greater transparency and flexibility [31, 35]. These approaches support a more collaborative model of interaction, where the assistant becomes a configurable partner rather than a fixed service.

Evaluation Method: To evaluate these systems and interaction designs, intelligent assistant research typically draws on four methodological approaches: *in-the-wild studies* [1, 5, 17], *customized prototypes* [31, 50], *interviews* [11, 14], *storyboards* [40, 50, 56], and *Wizard-of-Oz (WoZ) simulations* [8, 9, 35]. In-the-wild deployments provide insights into real-world usage and long-term adaptation but are resource-intensive and difficult to scale. Customized prototypes, often built on top of commercial platforms, enable the testing of new interaction strategies but require sustained development and longitudinal testing. Interviews, surveys, and storyboards allow for efficient elicitation of user feedback in early-stage design, but they lack contextual grounding and real-time user reactions. WoZ simulations enable researchers to mimic assistant intelligence without requiring full implementation, offering a middle ground for exploring speculative features and interaction dynamics.

2.2 Simulation Platforms and Datasets in Designing Intelligent Assistants

Simulation platforms such as SIMACT [6] and OpenSHS [15] have contributed to smart environment modeling and human activity dataset generation for intelligent assistant. Habitat [38, 42, 47] and AI2-THOR [26] extend this line of work by providing scalable, photorealistic 3D environments that support embodied agents in navigation, interaction, and manipulation tasks. VirtualHome [39] complements these platforms by introducing a programmatic representation of household activities, where agents execute sequences of atomic actions derived from natural language to simulate complex daily tasks. Despite their advances, existing platfomrs face several limitations. While systems such as VirtualSmartHome and OpenSHS aim to generate labeled datasets for machine learning, they are still bound by fixed scenarios and handcrafted rules, limiting their ability to simulate emergent or evolving behaviors. Scripted simulations (e.g., SIMACT, VirtualSmartHome) provide reproducibility but restrict behavioral variability and adaptability, which are essential for studying dynamic assistive systems. Avatar-controlled platforms (e.g., IE Sim, OpenSHS) enable more naturalistic behavior but are labor-intensive and difficult to scale. Most platforms do not incorporate cognitive processes such as goal reasoning, memory decay, or error-prone decision-making. MASSHA [25] introduces a belief-desire-intention (BDI) reasoning model but does not support dynamic learning from interaction. These limitations highlight the need for simulation frameworks that support scalable, autonomous agents capable of generating human-like activities and responses to evolving contexts and interactions.

2.3 Large Language Model-based Agents

Large language model (LLM)-based agents have emerged as powerful tools for simulating human cognition and behavior across a wide range of domains, including cognitive modeling, human–computer interaction, social simulation, autonomous planning, and multi-agent collaboration [16, 22, 30, 34, 36, 45, 48, 52, 54, 55]. These agents demonstrate capabilities such as multi-step reasoning, goal-directed decision-making, dynamic preference adjustment, and coherent logic flow in complex environments [21, 43, 58]. Prior studies show that LLMs can simulate strategic responses to evolving scenarios and exhibit behaviors aligned with human-like planning, moral reasoning, and self-reflection [10]. Expanding on these cognitive capacities, LLM-based simulations have been used to generate conversational data for social science analysis and to explore social patterns and group behavior dynamics [4, 51, 59]. Park et al.[37], presented a multi-agent framework that replicates human behavior in domains such as policymaking and social interaction, grounded in the Big Five personality traits [24]. More works further explored the integration of psychological constructs—such as Maslow's hierarchy of needs [32] and the Ten-Item Personality Inventory (TIPI)[19]—to personalize agent behavior [7, 29]. From a cognitive modeling perspective, recent systems have incorporated emotion, memory, and social values to enrich agent behavior, while others dynamically adapt to changing environmental and physical contexts [49]. Collectively, these advancements support the development of LLM-based agents capable of simulating interactive, adaptive, and human-aligned behavior. Building on the success of LLM-based agents, our work focuses on developing a simulation framework to enable efficient and automated research on human–assistant interactions.

3 Background and Motivation

3.1 Human–Assistant Interaction Research Workflow and Limitations

Human–assistant interaction research typically requires long-term experimentation and extended data collection periods. Fig. 1(a) illustrates a typical workflow, which consists of stages from study planning to data consolidation.

(1) The process begins with study planning, where researchers define research goals based on user needs (e.g., providing timely reminders for scheduled events) and formulate specific research questions (e.g., "Can proactive voice assistants improve user punctuality?").
(2) In the system design phase, researchers develop evaluation metrics and plan the interaction mechanisms necessary to address their research questions. Common evaluation methods include Likert-scale ratings, task completion rates, and qualitative analyses of user feedback.
(3) During prototype development, researchers operate the study through interactive systems or simulated interactions, employing methods such as prototyping, interviews, storyboards, and Wizard-of-Oz (WoZ) simulations. While prototyping enables exploration of novel designs, it can be resource-intensive, difficult to modify after deployment, and sometimes fails to elicit authentic engagement when using static or scripted approaches.

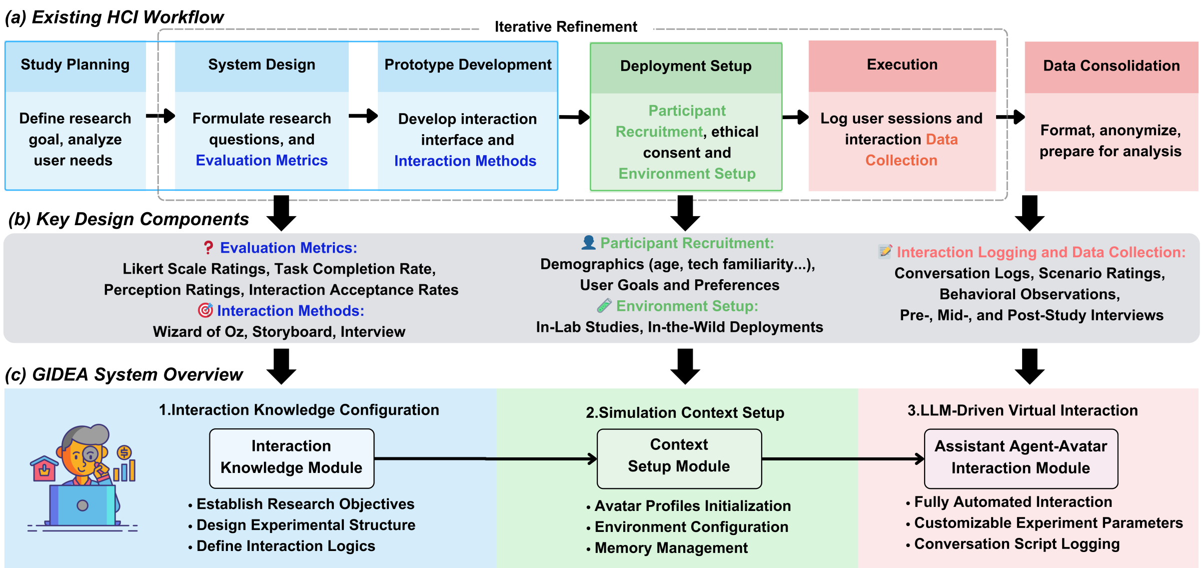


Fig. 1. (a) Existing HCI studies follow a resource-intensive workflow from design to deployment and data collection. (b) Core components span evaluation metrics, assistant interfaces, participant traits, and physical environments. (c) The proposed LLM-driven framework enables efficient simulation through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

(4) The deployment setup phase involves obtaining ethical approval, recruiting participants, and configuring the physical environment for data collection, often requiring manual installation of sensors or smart devices. Participant recruitment is often time-consuming and subject to privacy constraints, particularly when targeting diverse demographic groups.
(5) The execution phase carries out the study, either in controlled laboratory settings or in-the-wild environments. Each setting presents trade-offs: laboratory studies may limit ecological validity, while field deployments introduce challenges in environmental variability and long-term engagement.
(6) Finally, the data consolidation phase formats, anonymizes, and prepares collected interaction logs for subsequent analysis. Manual transcription and coding are common but labor-intensive, potentially introducing inconsistencies that affect traceability.

Fig. 1(b) further summarizes key designable components across the workflow, including evaluation metrics, interaction methods, participant recruitment criteria, environmental setup strategies, and data collection methods. These components are critical to shaping study outcomes but also contribute to the overall complexity and resource demands of the research process. Our platform focuses on these components and leverages generative AI to support human–assistant interaction research.

3.2 Opportunities with Generative AI

We leverage the social simulation capabilities of generative AI to provide an alternative to existing human–assistant interaction experiments. Building on the conventional workflow summarized in Fig. 1(a), our simulation framework (Fig. 1(c)) mirrors the structure of prior research while introducing generative components to replace or augment core stages.

Specifically, large language models (LLMs) are integrated into the assistant–avatar interaction phase of the workflow. Foundational study parameters—such as research objectives, study methods, evaluation metrics, and the interaction logic between the assistant agent and avatars—are manually specified through a structured interaction

Knowledge configuration (Fig.1(c).1), providing the backbone for replicating prior studies or designing new experiments. In the simulation context setup phase (Fig.1(c).2), researchers prepare avatar personality profiles, environmental configurations, and memory initialization, which can be manually defined or optionally generated via LLMs. Once the setup is complete, the LLM-driven virtual interaction module (Fig. 1(c).3) autonomously simulates assistant–avatar dialogues. At this stage, generative models produce coherent, context-aware, multi-turn interactions that emulate realistic user responses and assistant behaviors, effectively virtualizing the data collection phase of human-assistant interaction research.

We develop GIDEA as a modular simulation framework that integrates these components into a cohesive system. By aligning each simulation stage with its counterpart in existing experimentation, GIDEA enables researchers to systematically study intelligent assistant behaviors, compare interaction strategies, and iterate on experimental designs without the resource constraints of physical deployments. This framework provides a principled foundation for scalable, flexible, and traceable human–assistant interaction research.

4 System Design

4.1 Design Goals

GIDEA is designed to support scalable, flexible, and reproducible simulation-based human–assistant interaction research. Our design is guided by the following five key goals:

- **Seamless configuration of user, environment, and assistant models.** Studies differ in user traits, interaction methods, environmental contexts and evaluation metrics. GIDEA enables modular reconfiguration of personas, environmental contexts, interaction policies, and evaluation instruments, allowing researchers to flexibly adapt simulations to different study designs without modifying system internals.
- **Scenario-based study design through reusable templates.** Many studies follow structured routines, such as morning preparations, proactive interruptions, or task reminders. GIDEA abstracts these routines into modular templates with predefined roles, triggers, and actions. Researchers can define targeted interaction scenarios and easily customize them by adjusting conditions, actions, or conversational flows without reimplementing core simulation logic.
- **Simulation of dynamic temporal and contextual changes.** Real-world human-assistant interactions are triggered by evolving contexts over time. GIDEA models temporal progressions (e.g., time of day changes) and environmental state shifts (e.g., location, attention, device status), enabling agents to operate within dynamic, realistic environments during simulation runs.
- **Baseline human-aligned behavioral modeling.** To promote coherent interaction patterns, GIDEA assigns each simulated avatar a personality profile based on Ten-Item Personality Inventory (TIPI) [19] scores and defined narrative traits. While the current implementation uses lightweight, static profiles, the modular architecture is designed to support integration with external cognitive, affective, or memory-based behavioral models in future extensions.
- **Traceability of simulation workflows.** GIDEA logs experimental parameters, environmental contexts, and assistant agent–avatar interaction histories in structured textual formats. These detailed records support transparent post-hoc analysis, enable reconstruction of study setups, and facilitate cross-comparison across different simulation runs, even when model outputs vary due to inherent stochasticity.

4.2 Module Descriptions

GIDEA operates through three coordinated modules that together support the design, configuration, and execution of simulated human–assistant interactions: the *Interaction Knowledge Module*, which encodes the study design and logic; the *Context Setup Module*, which defines avatar and environment configurations; and the *Assistant Agent–Avatar Interaction Module*, which governs runtime interaction dynamics (Fig. 1(c)). In the following paragraphs, we present the overall design of each module

4.2.1 Interaction Knowledge Module. This module defines the structure and logic of each simulation and is initialized at the beginning of each case study (An example of interaction knowledge is shown in Appendix A.1).
Purpose: It serves as the foundation of each simulation, capturing the experimental structure and assistant agent–avatar interaction logic that define a given simulation instance.
Configuration: Researchers begin by defining their study goals and then configure the module accordingly. They populate it with study-relevant metadata, such as evaluation metrics, which may be qualitative (e.g., pre-, mid-, and post-study interview questions) or quantitative (e.g., Likert-scale ratings, task completion rates)—all commonly used in human–assistant interaction studies. Depending on the research scenario, the assistant agent may act as a study facilitator or proactive assistant, while the avatar assumes the role of a simulated participant.
Role Separation: To preserve realism and prevent knowledge leakage, the assistant agent and avatar are configured with asymmetric knowledge contexts. The assistant agent has access to the full case study setup, including the study instructions, configured evaluation metrics, and predefined logic for how it should behave during the simulation. In contrast, avatars are restricted to participant-facing information, such as environmental context and interaction history.
Implementation: The full simulation specification is encoded in a structured JSON schema, distilled from prior studies and interviews with HCI researchers. This schema supports sanity checking, reproducibility, and consistent replication across studies. The schema has been applied and evaluated across ten case studies.

4.2.2 Context Setup Module. This module prepares the dynamic elements required for each simulation run, including avatar profiles, environment states, and initial memory.
Avatar Profiles: Avatars are constructed using demographic data, psychological traits, and contextual attributes based on the original study population. Each avatar is assigned a personality profile based on the Ten-Item Personality Inventory (TIPI) [19], which enriches their behavioral modeling. Note that if a study provides TIPI scores, we can also match the simulated TIPI with the original scores.
Persona Narratives: The assigned personality traits, combined with the demographic and other relevant attributes, are used to generate detailed background narratives for each avatar—such as lifestyle routines, communication styles, and preferences—to support lifelike behavior during interaction (see Appendix A.2 for persona narrative examples).
Environment Configuration: The physical environment is initialized to reflect the study setting (e.g., home, office). Rather than including all visible objects, the system selects only those likely to emerge during interaction, supporting clarity and task relevance. These curated interactive objects are accessible to both assistant agent and avatar and serve as cues for context-aware interactions (see Appendix A.2 for an environment example).
Memory Management: All contextual elements—including avatar profiles and environment configurations—are initialized during the first iteration and reloaded in each subsequent simulation round. This supports coherent, persona-aligned responses during assistant agent and avatar interactions. Since all case studies are conducted within a short time window (typically within a few hours), long-term memory management is not implemented.

4.2.3 Assistant Agent–Avatar Interaction Module. This module governs the runtime execution of the simulation by managing the dialogue and decision-making processes between the assistant agent and avatars.
Interaction Dynamics: Both assistant agent and avatars are powered by GPT-4o. They operate independently, generating responses based on their individual prompts, task goals, personas, and accessible context (see Appendix A.4). Interaction histories, memory states, and environmental changes are systematically logged to support traceable analysis and reproducibility.

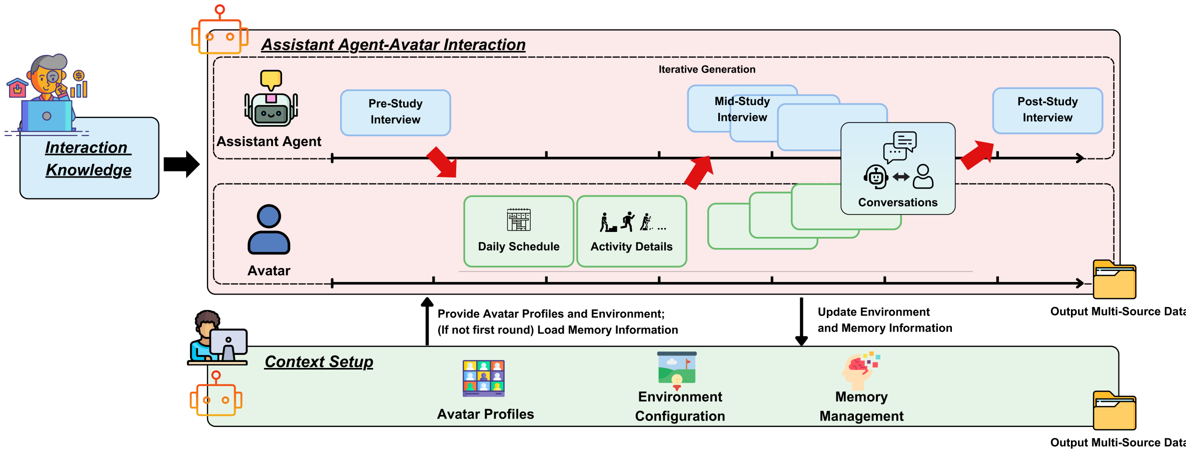


Fig. 2. Simulation workflow of GIDEA. Researchers specify study goals and interaction parameters via the Interaction Knowledge module, while avatar profiles and environmental states are configured through the Context Setup module. During runtime, the Assistant–Avatar Interaction module dynamically executes study iterations, generates proactive dialogues, and updates memory and environment states across simulation rounds.

Knowledge Isolation: The assistant agent and avatars do not share internal states or reasoning processes. The assistant agent cannot access the avatar's internal state unless made explicit during interaction. For example, the assistant agent cannot reference or infer the avatar's unspoken intentions or internal states unless revealed during the conversation. Because these interactions are generated dynamically using large language models, we implement strategies to reduce factual inconsistencies or hallucinations. Each response is constrained by (1) Memory snapshots that include shared interaction history and role-specific internal reasoning for the assistant agent and avatar, (2) Real-time environmental context derived from the simulation state, and (3) Role-specific knowledge boundaries that enforce strict separation between the assistant and avatar's accessible information.

More advanced techniques such as retrieval augmentation [20, 28] or factual consistency checking [23, 27] can be applied to further reduce potential hallucination. However, we did not observe significant hallucinations with our current strategy, so we chose not to incorporate them in this version.

4.3 Simulation Workflow

Fig. 2 illustrates the execution flow of GIDEA. The simulation proceeds through a structured sequence of phases designed to support realistic, traceable human–assistant interactions. Below, we describe the key steps that take place during each simulation run. A recorded demo based on the platform was also submitted.

1. Initialization Phase. The system begins by loading the interaction knowledge files and configuring the simulation context. This includes avatar profiles, environment layout, and prior event history to ensure consistency and completeness.

2. Pre- and Mid- Study Interviews (Optional). If defined by the study design, the assistant conducts a pre-study or mid-study interview with the avatar. The avatar's responses are logged and later used for post-simulation evaluation and comparison.

3. Schedule Generation. Human–assistant interactions are often grounded in daily routines. The system generates a basic, time-stamped daily schedule for each avatar using a GPT-4o prompt (see Appendix A.3). The prompt is conditioned on the avatar's profile, environment, and memory state. Typical activities might include breakfast preparation, checking notifications, or starting work.

4. Activity Enrichment. Each scheduled activity is expanded with additional context using another prompt (see Appendix A.3). The enrichment includes motivations, subtasks, expected environmental interactions, and potential challenges. This enrichment process supports fine-grained avatar behavior modeling and provides the assistant agent with context for initiating targeted interactions, enabling more realistic interaction details and capturing potential opportunities for timely or context-relevant interaction.

Example: For a base activity such as "Alex starts working on a new art project at the table in the main room," the system generates a detailed scenario in which Alex draws inspiration from a recent trip and adjusts the lighting to enhance his workspace. This elaboration reveals his motivation (creative expression inspired by ocean memories), subtasks (preparing paints, adjusting blinds, evaluating progress), environmental interactions (engaging with sunlight, paint tools, hydration), and potential challenges (distraction, time unawareness).

5. Assistant Agent–Avatar Interaction. Based on the enriched context, the assistant may engage with the avatar either proactively or responsively, depending on the scenario logic. Both roles are powered by separate GPT-4o instances (see examples in Appendix A.4). They operate independently, with shared memory (e.g., conversation history) and non-shared attributes (e.g., goals, role-based knowledge). This asymmetry mimics real-world facilitator–participant dynamics.

6. State Update. After each interaction, the system updates the avatar's memory and the environmental state to reflect what happened. These updates carry over into the next simulation round, allowing the simulation to maintain continuity and coherence.

7. Post-Study Interview. At the end of the simulation, the assistant may conduct a post-study interview to assess avatar's perspectives and the assistant agent's performance. This mirrors real-world methods used in evaluations of human–assistant interaction research.

8. Logging and Output. All simulation output—including multi-source data such as schedules, activity details, dialogue transcripts, environmental states, and memory logs—are saved in structured text formats. This supports reproducibility and facilitates analysis.

5 System Evaluation

5.1 Evaluation Methodology

Evaluating the realism of human decision-making in human–assistant interactions remains a challenge due to the limited availability of publicly accessible datasets that capture both observable behaviors and internal contextual states. We assess GIDEA's effectiveness by examining whether it can replicate the key findings or observations from prior human-participated studies, i.e., a replication-based evaluation strategy. To find relevant studies, we began by surveying the HCI literature from major venues such as IMWUT and CHI using the keywords "voice assistant," "conversational agent," and "proactive system." The studies were selected following these criteria:

- The interaction must involve human participants and an assistant agent executing a clearly defined task. The assistant agent's behavior must remain consistent across all participants and may be instantiated via a system prototype or a human confederate in a Wizard-of-Oz configuration.
- The studies must be well-documented for possible replication. All interaction topics or tasks must be accessible to participants without requiring specialized domain knowledge.
- The interaction must be conducted via or convertible to text, as the current version of the simulation environment supports only textual modalities for both input and output.
- The case studies reflect diverse research focus or themes on human-assistant interactions.

Following these criteria, we identified ten case studies [8, 9, 11, 14, 31, 35, 40, 50, 56, 57] that reflect four key themes in assistant research, including *Personalization and Social Framing*, *Proactivity and Context-Awareness*, *Managing Attention and Interruptibility*, and *User Control and In-Situ Configuration*. Table 1 provides an overview of their research objectives, study method, participant and physical setups, and evaluation metrics.

Table 1. Summary of case studies, research objectives, methods, setup, and evaluation metrics.

Theme	Case Study (CS) and Its Research Objectives	Study Method	Participant and Physical Setup	Evaluation Metrics
Personalization and Social Framing	CS1 [56]: Users' expectations for assistant personality and personalization	Storyboard-based interviews	15 participants, tech familiarity; 10 scenario-based storyboards	Assessment of personality score; Qualitative analysis of assistant personality
	CS2 [11]: Comparing human-human and human-agent conversation expectations and value	In-lab, semi-structured interviews	17 participants, tech familiarity; Smart environment with smart speaker	Qualitative analysis of conversation characteristics
	CS3 [9]: Effects pf conversational style perceived intelligence, trust, and metaphorical framing in older adults	In-home study, Wizard-of-Oz with interviews	58 participants, later adulthood; Lab environment with smart speaker	Likert-scale acceptability ratings; Qualitative analysis of metaphor usage
Proactivity and Context-Awareness	CS4 [14]: Effects of proactive feedback styles on user perception and reflection	In-lab study, Wizard-of-Oz study with interviews	30 participants; Lab environment with smart speaker	Perception ratings (e.g., helpfulness, appropriateness); Qualitative analysis on proactive persuasion
	CS5 [50]: Influence of social presence and co-activity on proactive assistant acceptance	In-home study using prototype	18 participants, tech familiar; Smart environment with smart speaker	Perception ratings (e.g., comfort, trust, disruption); Qualitative analysis of proactive initiations
	CS6 [57]: User preferences for when and how proactive behaviors should be initiated	Storyboard-based interviews	15 participants, tech familiarity; 9 scenario-based storyboards	Scenario ranking; Qualitative analysis of design tensions (e.g., timing, agency)
Managing Attention and Interruptibility	CS7 [8]: Personal, physical, and social contexts affecting interruptibility	In-home study, Wizard-of-Oz with interviews	40 participants; Smart environment with smart speaker	Interaction acceptance rate; Qualitative analysis of interruption timing
	CS8 [40]: User evaluations of proactive assistant interactions across utility, tone, and privacy	Storyboard-based interviews	47 participants; 8 scenario-based storyboards	Scenario ratings; Qualitative analysis of interaction acceptance
User Control and In-Situ Configuration	CS9 [35]: Timing and alignment of assistant actions with user agency in proactive support	In-lab study, Wizard-of-Oz study with interviews	12 participants; Smart environment with smart speaker	Interaction acceptance rate; Qualitative analysis of interruption timing
	CS10 [31]: Interaction patterns and behavior modeling during in-situ rule configuration	In-lab study, Wizard-of-Oz study with interviews	16 participants; Smart environment with smart speaker	Interaction flow proportions; Qualitative analysis of rule-setting preferences

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. Rather than attempting to reproduce participant responses verbatim, we assess whether the simulation captures comparable interaction patterns and assistant-driven effects. We employ two complementary evaluation methods:

(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–avatar interactions, 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 that are conceptually consistent with those derived from human participants.

(2) Interaction Log Analysis: To contextualize the semantic similarity results, we analyze assistant–avatar interaction logs through both qualitative and quantitative lenses:

- *Quantitative patterns:* We conduct statistical analysis using a range of methods to evaluate behavioral metrics, going beyond distribution comparisons to include visualizations and summary statistics aligned with those used in the original studies.
- *Qualitative evidence:* Representative dialogues and reasoning traces are extracted to illustrate parallels with the original study's reported behaviors or explanations.

Together, these methods allow us to evaluate the fidelity of simulation outcomes in terms of both high-level semantic alignment and detailed behavioral consistency. This framework provides a robust and interpretable basis for comparing simulated and empirical results across diverse study designs and reporting formats.

5.2 Simulation Configuration

To replicate each of the 10 selected case studies, we configured GIDEA to mirror the original experimental settings as closely as possible. This involved translating study-specific components into simulation modules, including user personas, assistant behaviors, environmental layouts, and timing structures.

User Personas: Each case study targeted specific participant groups-such as older adults, students, or tech novices-to elicit feedback aligned with its research objectives. To simulate these populations, we generate avatar profiles using structured profile representations. These profiles include demographic attributes (e.g., age, occupation) and other relevant attributes (e.g., technology familiarity, household type) that are sampled based on a given distributions (e.g., the reported ones in the original studies). Each avatar is also assigned a TIPI-based personality score and a narrative description reflecting the intended characteristics (detailed in system design).

Environment Setup: By default, simulation initialize within a one-bedroom indoor smart home layout populated with interactive objects, including appliances (e.g., floor lamp, air conditioner, fan), tools (e.g., vacuum, remote control), and furniture (e.g., sofa, table). Researchers can optionally reconfigured the space to simulate different living arrangements, such as shared-living environment or a single-family house, depending on the research study requirements. This flexibility allows case studies to reflect diverse domestic contexts, which may influence interaction patterns.

Scenario Modeling: Several case studies included example scenarios to guide participant interactions, such as sample dialogues, storyboard narratives, or descriptions of desired experimental scenes. When these scenarios were provided in textual form, they were directly incorporated into the simulation instructions. For non-textual formats, we translated the content into narrative-driven textual instructions. In particular, studies employing storyboards were converted into structured narratives that preserved the original interaction flow and conversations.

Assistant Agent and Avatar Behavior Policies: Both the assistant agent and the avatar were implemented using a large language model (LLM), enabling them to generate naturalistic, role-appropriate behaviors. Assistant agent behaviors were guided by its designs and task objectives described in the original user studies. Depending on the study scenario, the assistant agent was assigned distinct roles to match the requirements of each study scenario—for example, serving as proactive recommender, experiment facilitators, or context-aware interactor. The avatar, in turn, was configured to simulate a human participant—responding to assistant prompts, expressing preferences, making decisions, and reflecting on experiences as appropriate for the study context.

Treating both the assistant agent and avatar as role-playing participants, we provided detailed, role-specific instructions outlining their expected behaviors and responsibilities. For studies involving multi-turn interactions, both the assistant agent and avatar were equipped with conversation memory to retain prior exchanges and ensure interaction coherence. The assistant agent was configured to support either single-turn or multi-turn dialogue, depending on the study's requirements. The number of interaction rounds was customizable to replicate the original experimental protocol. These behavior policies governed when and how the assistant initiated actions, provided suggestions, or responded to avatar input in a manner consistent with the study's design.

Evaluation Instruments: Some user studies incorporate pre-, mid-, or post-study questionnaires and interviews to collect user feedback. In our simulated experiments, we replicated these instruments by posing the same questions to either the avatar or the assistant agent at the corresponding points in the interaction timeline. This approach allows us to elicit realistic, context-sensitive feedback based on the avatar's current memory and interaction history, thereby preserving the temporal alignment and evaluative intent of the original study.

5.3 Aggregate Replication Accuracy Across Case Studies

This part presents the replication accuracy across 10 case studies. To fairly compare the semantic similarity between the original study and generated simulated results on each research questions, we summarized (1) the extracted original findings and (2) the simulated interaction data and questionnaire feedback for each RQ, following the rule described below:

- **Original Study Summary:** For original case study findings, we extract the directly corresponding content from the original case study paper for each RQ (e.g., from the evaluation, discussion or findings sections). We then used GPT-4o to generate focused summaries based on the content, ensuring the output follows a question and answer format (example in Appendix A.5).
- **GIDEA Simulation Summary:** GIDEA-generated assistant agent and avatar interaction logs do not include an organized discussion of findings. Each simulated study includes (1) multi-turn interaction logs between the assistant and avatar agents and (2) filled-out questionnaires that parallel the original study's evaluation. To extract comparable answers, we again used GPT-4o with the same summarization prompt as for the original answers, instructing it to synthesize a coherent response to the RQ based on observed behavioral patterns and questionnaire feedback (example in Appendix A.5).

This alignment in summarization method ensures that the original and simulated answers are directly comparable, despite differences in their source material. Furthermore, to minimize the risk that the original study content was memorized during the large language model's pretraining, we only exposed the model to high-level findings extracted manually from the papers rather than direct quotes or detailed textual material. Summarization was performed uniformly for both the original and simulated sources using the same prompts, ensuring that evaluation results reflect reasoning over summarized content rather than memory recall. Then we use the `all-mpnet-base-v2` model from Sentence-Transformers [41, 44] to compute cosine similarity between each pair of original and simulated RQ answers. The result is a similarity score for each research question. Fig. 3 visualizes the distribution of similarity scores across all 25 RQs. Each bar represents the similarity score for a single RQ, grouped and color-coded by case study. The average similarity per case study is also shown in the legend, offering a study-level view of replication quality.

To further visualize and analyze the replication quality of GIDEA, we grouped the results by (a) study theme and (b) experimental mode. As shown in Fig. 4a, studies focused on *Proactivity* and *Personalization* achieved the highest

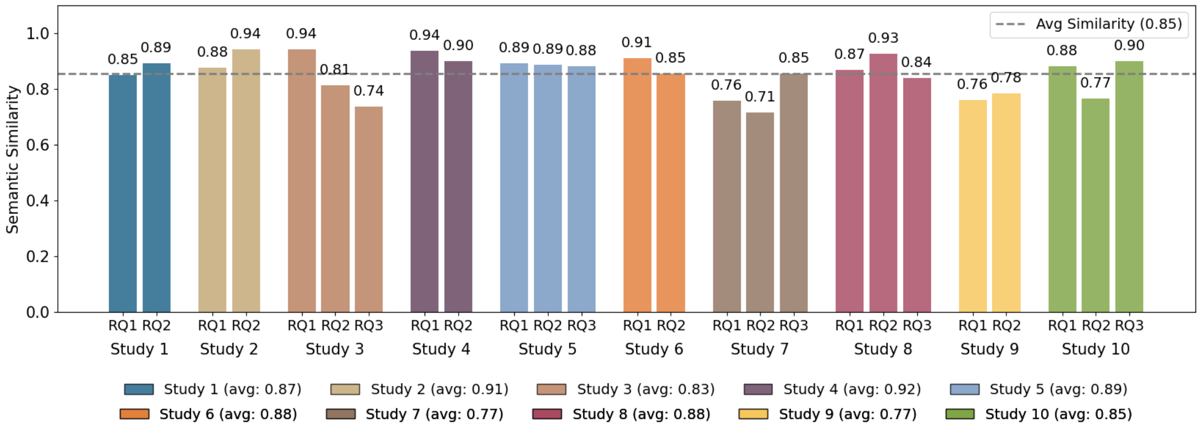
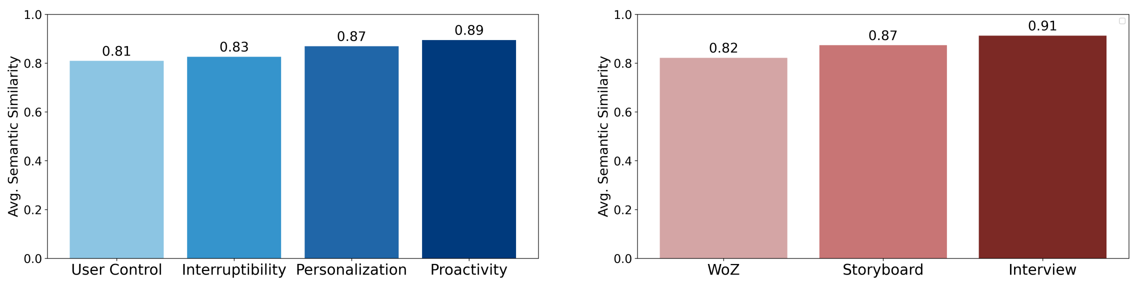


Fig. 3. Semantic similarity scores comparing simulated and original study responses for each research question (RQ) across 10 case studies. The red line shows the overall average similarity (0.88). Study-specific average similarities are in the legend.



(a) Average semantic similarity grouped by study theme. (b) Average semantic similarity grouped by study mode.

Fig. 4. Average semantic similarity scores grouped by (a) study theme—User Control, Interruptibility, Personalization, and Proactivity—and (b) study mode—Wizard-of-Oz (WoZ), Storyboard, and Interview.

semantic alignment (0.89 and 0.87, respectively), while *User Control* studies yielded the lowest average similarity (0.81). We observed that study theme was associated with differences in semantic similarity, prompting further analysis based on study mode to explore potential underlying causes. Notably, both User Control studies employed a Wizard-of-Oz (WoZ) methodology, which often introduces variability due to dynamic and spontaneous human activities. As a result, reactions and responses triggered by activities only partially aligned with the original case study findings. Moreover, the original case studies did not provide detailed behavioral data, limiting our ability to adjust the simulation to better match the original conditions. We further analyzed the results by study mode. As shown in Fig. 4b, simulations of *Interview*-based studies achieved the highest average similarity (0.91), followed by *Storyboard* (0.87), and *WoZ* (0.82). Interview and storyboard studies typically provide structured questionnaires or detailed scenario examples, offering clear rationales for user responses and reducing ambiguity during simulation. This explains why they achieve higher semantic similarity scores compared to WoZ-based studies. Overall, these findings confirm that our simulation framework achieves a high semantic similarity in reproducing study findings, with variations closely tied to study design and some misalignment occurring when more spontaneous factors are involved.

Table 2. 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"

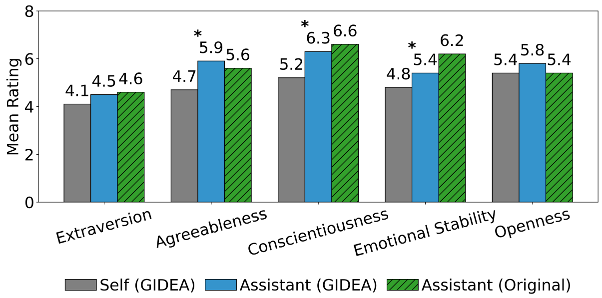
5.4 Case Study-Level Evaluation Results

To better interpret GIDEA's effectiveness across diverse research contexts, we conducted a study-level analysis based on the full conversational transcripts between the assistant agent and avatar agents. For each case study, we report both qualitative observations and quantitative patterns extracted from the interaction logs,highlighting specific replication outcomes and discussing unique factors that influenced simulation quality. Before presenting case-by-case results, we summarize several key insights derived from the ten replicated studies. Each insight is supported by specific examples, which are elaborated in the corresponding case study sections below.

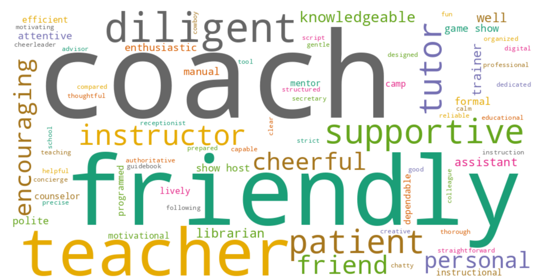
- **Preservation of High-Level Behavioral Patterns**: GIDEA preserved core behavioral patterns observed in original human–assistant interaction experiments across the ten replicated studies, such as activity structures, conversational framing, proactive interaction timing, and smart home rule programming strategies.
- **Alignment of Thematic Reasoning**: GIDEA avatars demonstrated reasoning that paralleled human participants. Both quantitative outcomes (e.g., conversation response rates, scenario rankings) and qualitative rationales (e.g., desire for autonomy, sensitivity to timing, comfort with emotional tone) showed strong thematic alignment, making it suitable for simulating user-centered interaction studies.
- **Context-Sensitive Interaction Modeling**: GIDEA reflected users' tendency to adjust engagement based on contextual factors such as physical activity state, emotional readiness, and privacy concerns, reinforcing its utility for studying experiments that include contextual information as a critical factor.
- **Strong Modeling of Rule-Based Configuration Behavior**: In the in-situ programming studies, GIDEA mirrored users' programming approaches, such as starting with direct control commands and progressively refining rules based on spatial or environmental cues, demonstrating its ability to replicate smart home interaction and collaboration-related studies.
- **Potential for Group-Level Reflection Simulation**: Our evaluation suggests that GIDEA is capable of generating human-like reflections across diverse avatars, making it promising for applications that require capturing group-level opinions, participatory design feedback, or early-stage concept evaluations.

5.4.1 Theme: Personalization and Social Framing. This group of studies investigates how personalization and social framing influence user perceptions of home assistants. Despite differences in focus, all three studies assign agents human-like traits—such as personality profiles (e.g., TIPI scoring), conversational style, and tone. They tend to use storyboard or Wizard-of-Oz (WoZ) methods to explore user acceptance of more socially attuned, personalized agents, due to current technological limitations in prototype design.

- **Case Study 1 - Personalization Preferences**



(a) Case Study 1: Mean personality trait ratings comparing users' self-perception and their imagined assistant (agent), across simulated and original datasets. * Significant difference between Self and Assistant ratings in GIDEA (p < 0.01)



(b) Case Study 3: Word cloud summarizing metaphorical descriptions of RAVA across all conditions.

Fig. 5. Personalization and Social Framing.

Table 3. Case Study 2: Characteristics of Meaningful Conversation in Human vs. Agent Interaction

Theme	Simulated Avatar Quote	Human Participant Quote
Emotional Connection and Support	"It was emotionally grounding, providing a sense of connection and **mutual support** that left me feeling more at peace."	"Social conversations in my daily life often revolve around sharing personal experiences and reflecting on emotional or mental states. They foster a sense of connection and provide **emotional support**."
Functional Trust in Agents	"I appreciate that the assistant **remembers my preferences** and does not make me repeat myself."	"Agent trust was framed as **system reliability and memory**, not emotional safety or shared vulnerability."
Role of Humor in Interaction	"The assistant's joke lightened the mood, but I **still want clear answers**."	"Human humor fostered bonding; AI humor was seen as a novelty that must **not interfere with task clarity**."

- **Overview.** Zargham et al.[56] explored how participants imagine the ideal personality and customization features of home assistants using a storyboard-based experimental method. To replicate this case study, we initialized simulated avatars in GIDEA with the same mean personality trait scores as the original participants' self-assessments and exposed them to the same storyboard scenarios depicting daily interactions between home assistants and users.
- **Analysis.** As shown in Fig. 5a, 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. Simulated avatars in GIDEA also reflected participants' nuanced expectations for assistant behavior and interaction style. As shown in Table 2, 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.

- **Case Study 2 – Characteristics of Meaningful Conversation**
  - **Overview.** Clark et al. [11] explored what makes conversations meaningful by comparing human–human and human–assistant interactions. Through semi-structured interviews, they identified key conversational characteristics such as emotional connection, trust, mutual understanding, active listening, and humor. To replicate this study, we exposed the assistant agent to these key conversational characteristics and allowed it to freely lead interviews with avatars around these topics. This design choice was necessary because the original paper did not provide specific interview questions and discussed only high-level topics. Our evaluation focuses on qualitatively comparing quotes by directly matching avatar-generated quotes with participant quotes reported in the original study, rather than performing any statistical or structural conversation analysis, consistent with the analysis approach used in the original case study.
  - **Analysis.** Participants in the original study described meaningful conversations as emotionally supportive and grounded in personal exchange, often facilitated through empathy and attentiveness. These same themes emerged in GIDEA. Shown in Table 3, one avatar described the interaction as "emotionally grounding," highlighting a sense of mutual support and peace. This parallels human reflections on the value of emotional connection in everyday conversations. However, the interpretation of trust diverged between contexts. In Clark et al.'s study, trust in agents was framed functionally—as system reliability, memory, and responsiveness—rather than emotional vulnerability. Avatars in the simulation similarly prioritized efficiency and consistency, suggesting that emotional trust may be less salient or less easily expressed in agent-mediated interactions. The treatment of humor in both studies revealed a shared expectation: while humor can enhance social bonding and make interactions feel more natural, it must not interfere with task effectiveness. Both studies reveal a consistent finding: while human–human conversations prioritize emotional depth, empathy, and relational trust, conversations with artificial agents are judged primarily by their efficiency, reliability, and task utility. Yet, avatars in the simulation also demonstrated an emerging desire for socially intelligent agents that could balance these functional expectations with humanlike attentiveness. These findings suggest that as AI becomes more integrated into daily life, conversational agents must evolve not only to fulfill tasks but also to engage with users in ways that feel meaningful and emotionally aware.
- **Case Study 3 – Age, Conversational Style, and Metaphorical Perceptions**
  - **Overview.** Chin et al. [9] conducted a Wizard-of-Oz study to examine how conversational style (formal versus informal) shapes perceptions of virtual assistants (VAs) among older and middle-aged adults. To replicate this study, GIDEA simulated the interaction context using a *Pattern Recall Game*. Avatars

Table 4. Case Study 3: Age, Conversational Style, and Metaphorical Perceptions

Theme	Simulated Avatar Quote	Human Participant Quote
Perceived Intelligence and Trust from Formal Style	"RAVA's instructions were clear and easy to follow—I **trust** her judgment."	Formal assistants were described as more structured and intelligent, enhancing perceptions of **trust**.
Metaphorical Framing of Informal Assistants	"RAVA is like a cheerleader" / "a fun, supportive **friend**"	Informal assistants were metaphorically described as "**friend**," or "family member."
Metaphorical Framing of Formal Assistants	"RAVA is like a well-organized **instructional manual**" / "a coach"	Formal assistants were framed as "teacher," "**guidebook**," or "librarian."

representing both age groups were randomly assigned to interact with the assistant agent in GIDEA, which role-played either a formal or informal conversational style during the game. After completing the task, avatars provided feedback through surveys designed to parallel the original study's evaluation measures, based on their experience interacting with the assistant.

- **Analysis.** The simulation reproduced key trends observed in the original study. Formal assistants were perceived as clearer and more reliable, while informal assistants were viewed as more relatable and emotionally supportive. As shown in Table 4, these differences were reflected in how avatars described the assistant: informal styles prompted comparisons to peers and supporters, while formal styles were associated with instructional or professional roles. Although conversational style did not significantly affect perceptions of intelligence, trust, or likeability overall, older adults—particularly those with lower agreeableness—exhibited stronger behavioral intentions toward informal assistants. Metaphor analysis revealed that older adults frequently described informal assistants using humanlike metaphors (e.g., *friend*, *family member*) and formal assistants using professional metaphors (e.g., *teacher*, *librarian*), whereas middle-aged adults showed no strong variation based on style. Word cloud visualize avatar metaphors (Fig. 5b) illustrate the divergence in social framing across conditions. Although GIDEA closely replicates participant preferences and general metaphor themes, it shows limitations in capturing the diversity of metaphors observed in the original study. The original textual data included a wider range of metaphorical connections, such as object-like metaphors, which are largely absent from the simulated outputs. Additionally, the simulated avatars tended to repeat a narrower set of metaphors—such as *coach*, *teacher*, and *instructor*—indicating reduced creativity and variation compared to human participants.

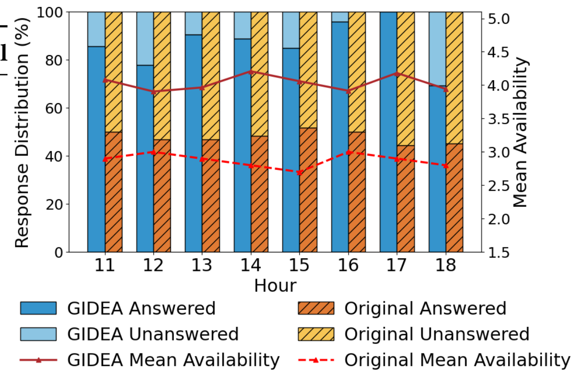
5.4.2 Theme: Proactivity and Context-Awareness. This group of case studies examines strategies for proactive conversation initiation, addressing the challenge that current agent designs often lack appropriate triggers to naturally start interactions with users. The studies investigate how to define valid initiation triggers and explore the use of multi-source data to support context-aware proactive communication.

- **Case Study 4 – Proactive Persuasion**

  - **Overview.** Dubiel et al. [14] investigated the impact of proactive feedback from a voice assistant on user perception and reflection in a voice-only food ordering scenario using a Wizard-of-Oz setup. To replicate this study, the assistant agent acted as the voice assistant, assisting the avatar in selecting a three-course meal and providing either solicited or unsolicited nutritional feedback. Avatars could choose to accept or ignore the advice, respond to active queries, and rate the assistant agent on social impression (trust, confidence, enthusiasm, and persuasiveness), appropriateness, menu choices, benevolence (interest,

Category	GIDEA	Original
Social Impression *Trust, Confidence, Enthusiasm, Persuasiveness*	5.5	4.8
Appropriateness *Behavior*	6.0	9.0
Menu Choices *Satisfaction*	7.0	3.5
Benevolence *Interest, Help, Preference*	4.0	3.5
Cognitive Load *Confidence, Tense, Calm, Concentration, Usage*	3.0	5.0

(a) Case Study 4: Median ratings between GIDEA and the original case study, aggregated across all subjects.



(b) Case Study 5: Hourly response distributions and mean availability between GIDEA and the original user studies.

Fig. 6. Proactivity and Context-Awareness

Table 5. Case Study 4: Proactive Persuasion and Perceived Agency

Theme	Simulated Avatar Quote	Human Participant Quote
Non-judgmental Tone and Pressure-Free Suggestions	"Human feedback might feel more personal but could **pressure more**—an assistant keeps choices pressure-free and neutral."	"If it were human, it would have been **more judgemental**, so it is better to receive suggestions from a robot." (P7)
Health Suggestions with Supportive Framing	"I prefer a voice assistant that offers **health-related advice** proactively as long as it's relevant and beneficial for my choices."	"It's a very new, creative and knowledgeable way to provide **insights about my health**."

help, and preference), and cognitive load (confidence, tense, calm, concentration) under both solicited and unsolicited interaction conditions.

- **Analysis.** Participants appreciated the assistant's impersonal and non-judgmental tone, which made suggestions feel less pressuring compared to feedback from a human. Fig. 6a presents a bar chart comparing the median ratings across categories, suggesting that GIDEA can replicate key patterns of user perception related to social impression, appropriateness, and cognitive impact. Ratings from the GIDEA simulation mirrored those from the original human-subject study to a certain degree. In addition to the quantitative alignment, qualitative reflections are summarized in Table 5 and achieved a high thematic similarity, highlighting parallel interpretations of tone, trust, and delivery style between the avatars and the original participants.

- **Case Study 5 – Proactive Initiation and Adaptive Feedback**
  - **Overview.** Wei et al. [50] conducted a three-week in-home deployment study to examine user perceptions of proactive smart speaker behavior using a Wizard-of-Oz setup. To replicate this study, avatars performed common home activities while the assistant agent acted as the smart speaker, initiating conversations at contextually appropriate moments to make the interactions feel natural and well-timed.

Table 6. Case Study 5: Proactive Initiation and Adaptive Feedback

Theme	Simulated Avatar Quote	Human Participant Quote
Opt-in Utterance Starters	"The assistant's gentle opening made it feel **optional**, which helped me stay in control while still being open to suggestions."	"I like the way the system starts the conversation... I can just say 'no' when it's inconvenient." "It gives me the **choice** to interact, so it performs more like humans."
Adaptive and Emotionally Responsive Feedback	"I want the assistant to learn about my preferences and tailor suggestions **based on my routine habits and mood**."	"After reporting a bad mood, participants expected dynamic follow-up like **mood-relevant suggestions**."

- **Analysis.** Fig. 6b shows a stacked bar and line chart comparing the hourly response distributions and mean availability between GIDEA and the original case studies. In the stacked bars, the darker color areas represent the proportion of responses that were answered within each hourly time period. We observe that the GIDEA simulation has a higher proportion of answered responses compared to the original study at most hours. This difference may reflect a limitation of the simulation. In this study setup, human activities were simulated; however, large language models have difficulty replicating the exact activity chains initiated from a human with limited demographic information. As a result, the scenarios encountered by the avatars can differ from those recorded in the original user study, for which detailed activity logs were not publicly available. Additionally, because GIDEA is powered by a generative AI model, it tends to actively generate answers for most queries, contributing to a higher answer rate compared to real human participants. The high availability scores and answer rates in GIDEA may therefore indicate some degree of misalignment in generating human-like activity patterns or potential misjudgments of contextual appropriateness. A fair comparison with the original study is limited, as no released dataset supports exact behavior chains with corresponding response rates. Despite these differences, thematic alignment is supported by Table 6. Avatars echoed the original finding that proactive prompts framed as optional fostered a sense of control. Similarly, under *Adaptive and Emotionally Responsive Feedback*, both avatars and participants emphasized the importance of context-sensitive suggestions, especially in relation to mood. The simulated agent's behavior aligned with participants' expectations for emotional intelligence, such as adjusting responses based on prior mood reports. This alignment across metrics and quotes supports the simulator's capacity to replicate complex, in-situ user preferences and social interaction norms.

- **Case Study 6 – Desirability and Delivery of Proactive Behavior**
  - **Overview.** Zargham et al. [57] investigated user perspectives on desirable proactive behaviors in voice assistants, specifically addressing the "proactivity dilemma"—how to intervene helpfully without being disruptive—using storyboards and interviews. Participants evaluated by ranking the nine assistant-initiated scenarios along three dimensions: usefulness, appropriateness, and invasiveness.
  - **Analysis.** Table 7 presents the quantitative comparison between GIDEA and the original study, showing the absolute differences in median rankings across scenarios. There is strong alignment in scenarios such as "Meeting Reminder" and "Emergency," where both original and simulated avatars agreed on high usefulness and low invasiveness. In contrast, larger discrepancies emerged in the scenarios "Disagreement Clarification" and "Health Risk." According to the avatars' conversation logs, for the "Health Risk" scenario, alerting users about health risks was considered important but could feel intrusive if not based on an

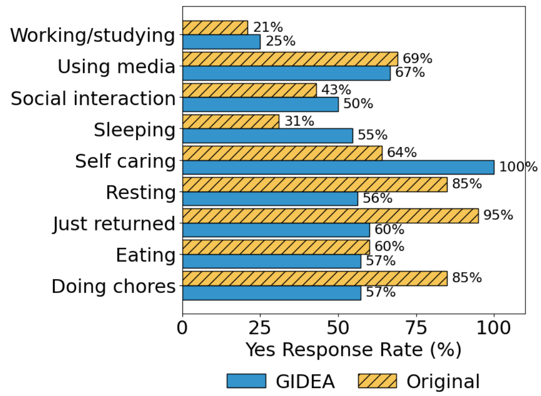
Table 7. Case Study 6: Median ranks from GIDEA and the original case study for usefulness, appropriateness, and invasiveness (1 = highest, 9 = lowest). Underlined scenarios indicate a ranking difference of 2 or more between GIDEA and the original study in **Usefulness** dimension.

Scenario	Useful		Appropriate		Invasive	
	GIDEA	Original	GIDEA	Original	GIDEA	Original
Emergency	1	1	1	1	9	9
Meeting Reminder	3	3	2	3	7	5
Technical Support	3	4	3	4	7	7
Health Risk	**4**	**2**	**5**	**4**	**3**	**5**
Cooking Inspiration	5	4	4	3	6	7
Disagreement Clarification	**5**	**8**	**4**	**8**	**6**	**1**
Nudging	7	6	6	6	3	4
Fact Checking	7	7	7	7	3	3
Fact Spoiler	9	9	9	9	1	3

Table 8. Case Study 6: Desirability and Delivery of Proactive Behavior

Theme	Simulated Avatar Quote	Human Participant Quote
Timely and Helpful Reminders	"I think it's great that the assistant can offer reminders when I'm running late; reminders can be **really helpful**."	"This is **a good feature** since [Jay] is making sure the user won't be late for her meeting."
Rejection vs. Irrelevance	"The suggestion wasn't wrong—it just **didn't fit what I needed in that moment**."	"Participants often declined suggestions not out of dislike, but because they **weren't necessary at the time**."
Transparency and Control	"I want the assistant to help, but I should be the one **deciding when and how**."	Participants emphasized the need for **permissions** and cues before the assistant acts.

immediate concern, leading to a lower usefulness ranking and a higher perception of invasiveness. Avatars initially appreciated the intervention, but repeated suggestions sometimes provoked annoyance, lowering the overall ranking of this scenario. Similarly, for "Disagreement Clarification," the original study found a direct quote that "this can be useful but it can hurt people's feelings when interrupting a natural conversation," leading to a lower usefulness ranking. Other than the discrepancies in these two scenarios, as summarized in Table 8, both simulated avatars and human participants valued timely and helpful reminders, distinguished between irrelevant and rejected suggestions, and emphasized the need for transparency and user control. These themes suggest that GIDEA avatars were able to reproduce not only scenario-level judgments but also the underlying rationale users expressed for accepting or rejecting proactive behavior. Together, these results highlight GIDEA's ability to replicate both aggregate behavioral ratings and fine-grained interaction expectations in proactive assistant design.



(a) Case Study 7: Comparison of yes-response rates between original and GIDEA studies across activity categories.

Scenario	GIDEA	Original
Alarm	4.0	4.0
Coughing	3.0	4.0
Tyre Change	4.0	4.0
Historical Fact	3.0	3.0
Time Clarification	3.5	3.5
Binge Watching	3.0	3.0
Setup Headphones	4.5	4.5
Quiz Spoiler	2.0	2.0

(b) Case Study 8: Comparison of median overall impression ratings (1=worst, 5=best) between original and GIDEA studies across proactive assistant scenarios.

Fig. 7. Managing Attention and Interruptibility

Table 9. Case Study 7: Opportune Timing for Proactive Interaction

Theme	Simulated Avatar Quote	Human Participant Quote
Low Interruptibility During High-Focus Tasks	"Any interruption before the end of my session could disrupt my workflow and reduce the effectiveness of **my study time**."	"I tended to not respond if I had made up my mind to focus on **studying**."
High Interruptibility During Transitions	"Since I have just finished a productive study session and I am in a **transition period**... I am open to any suggestions."	"[Participants] were more likely to be interruptible when they could temporarily **switch tasks** or during breaks."
Mood and Mental Readiness	"I'm **more open** when in a **relaxed** or open state of mind."	"**Negative states reduced willingness** to interact."
Sensitivity to Social Presence	"Although the smart assistant's timing is considerate, I **value my organic social interactions** with my roommate greatly."	"I was having a heated debate with my roommate, so it was **hard to interrupt** our conversation."

5.4.3 Theme: Managing Attention and Interruptibility. The studies in this group investigate strategies for managing user attention and interruptibility, exploring how proactive agents can use context cues—especially physical status changes—to initiate interactions at appropriate times.

- **Case Study 7 – Opportune Interaction Timing - Overview.** Cha et al. [8] investigated opportune moments for proactive interactions with smart speakers in domestic contexts through a one-week field study using a Wizard-of-Oz setup. To replicate this study, avatars performed common home activities while the assistant agent acted as the smart speaker, inquired about a conversation invitation, and calculated the yes-response rate.

- **Analysis.** The study examined how participants evaluate the timing of assistant-initiated interactions based on activity context. From Fig. 7a, we observed mismatched yes-response rates between the original study and the GIDEA simulation results. There was an extremely high response rate during activities such as sleeping and self-care, but a significantly lower response rate during other activities like resting, just returning home, and doing chores. We can summarize the pattern as follows: when participants were engaged in focus-required activities that demand high cognitive engagement—such as working or studying, using media, and participating in social interactions—the simulated results better captured the response rates based on the ongoing activity. In contrast, for less structured activities, more variability in decision-making was observed, suggesting that focus periods were not always visually detectable through activity type alone. Instead, internal cognitive or emotional states may be necessary to explain the decision to respond or not. From a thematic analysis perspective, as summarized in Table 9, we found strong alignment when explaining response rates at specific points, although this was not always fully reflected in the quantitative comparisons. Overall, we observed the limitation of GIDEA in precisely understanding when and how to intervene without compromising user agency or task focus. Nonetheless, its ability to reflect not only behavioral patterns but also situational reasoning confirms the simulator's fidelity in modeling human-centered interruptibility.

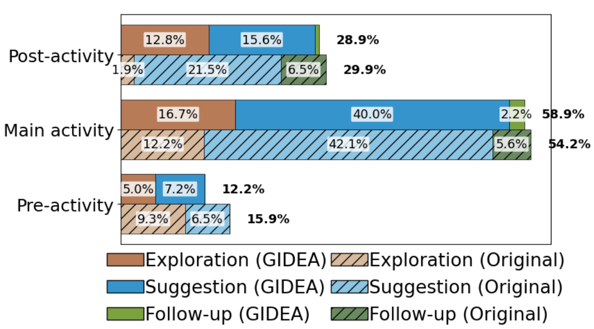
- **Case Study 8 – Perceived Interruptibility and Proactive Interaction Acceptance**
  - **Overview.** Reicherts et al. [40] examined user perceptions of proactive voice assistant behavior through an online study with a series of storyboards depicting voice assistant help users within different daily tasks.
  - **Analysis.** Avatar evaluated each interaction based on dimensions such as usefulness, pleasantness, appropriateness, and contextual sensitivity. As shown in Fig. 7b, scenario-level impression ratings in the simulation aligned with the original user judgments, with only minor variation across most conditions. Scenarios like "Alarm" and "Tyre Change," which were positively rated in the original dataset, also received high median scores from avatars. Quotes from simulated avatars and human participants in Table 10 further reinforce this alignment. Both original participants and simulated avatars emphasized the value of proactive features that were clearly beneficial and contextually appropriate. Comfort and acceptability were strongly tied to tone—supportive, calm phrasing contributed to positive impressions across both datasets. Additionally, shared concerns emerged regarding privacy and perceived intrusiveness. Despite technical reassurances in the original study, some participants expressed discomfort with the assistant's persistent presence. The simulation captured this skepticism, with avatars articulating unease when proactive behaviors felt overly invasive or unregulated. This consistency suggests that while GIDEA replicated surface-level preference patterns, it also captured the underlying rationale that shaped user receptivity.

5.4.4 Theme: User Control and In-Situ Configuration. The studies in this group examine human–agent collaboration in smart environments, emphasizing the importance of preserving user control and agency. They explore how proactive systems can offer evolving, personalized assistance while respecting users' preferences and autonomy.

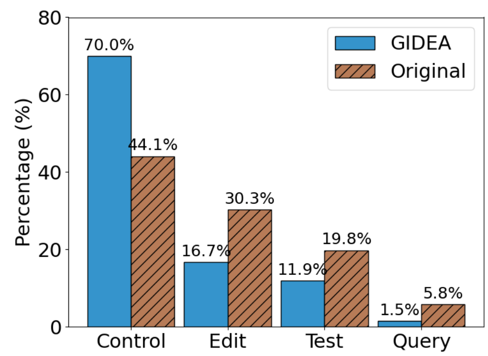
- **Case Study 9 – Respecting User Agency in Proactive VA Communication**
  - **Overview.** Oh et al. [35] explored how proactive voice assistants (VAs) in smart home environments can support user autonomy through communication strategies that balance initiative with sensitivity. Similar to other Wizard-of-Oz-based studies, avatars performed their daily home activities while the assistant agent acted as the smart speaker, asking if it was a good time to talk when detecting movements.
  - **Analysis.** Fig. 8a shows the proportion of post-, main-, and pre-activities, comparing the GIDEA simulation to the original study. Overall, the initiation of conversations and the use of different strategies align well with the original study across activity types. One discrepancy is that, during post-activity

Table 10. Case Study 8: Perceptions of Interruptibility and Proactive Interaction Acceptance

Theme	Simulated Avatar Quote	Human Participant Quote
Usefulness of Proactive Features	"This focuses on an essential aspect of the user's wellbeing, which can be **highly beneficial**."	"Overall, I found Jay's proactive behaviour **helpful**."
Tone and Appropriateness	"It feels like the assistant genuinely **understands my rhythm**. That's comforting."	Scenarios like S1 Alarm and S3 Tyre Change were positively received when delivered helpfully and **at appropriate moments**.
Privacy and Intrusiveness Concerns	"**Privacy concerns** could lead to discomfort or distrust among users. The approach might come off as intrusive rather than supportive."	"Jay appears to **always be listening**, but does Jay ever say 'Please turn off the microphone...'?"



(a) Case Study 9: Comparison of proactive assistant strategies by timing between GIDEA and the original user studies. The stacked bars show the distribution of strategies during three phases.



(b) Case Study 10: Comparison of interaction type distributions between GIDEA and the original user studies. Bars represent the percentage of interactions categorized as Control, Edit, Test, and Query.

Fig. 8. User Control and In-Situ Configuration

phases, the use of exploration strategies—where the assistant agent inquires about additional information, such as the avatar's preferences—is higher in the simulation. This likely reflects variation in human decision-making within similar activities. All conversations were initiated by human operators, and the operator's style may have influenced interaction timing, although limited information is available about it. Despite this discrepancy, GIDEA performed well in interpreting context to initiate conversations at appropriate times. The results support that proactive suggestions and follow-ups were more frequently accepted during or after main activities, aligning with natural task transitions. The simulation mirrored this temporal sensitivity, showing that avatars adjusted their responsiveness based on timing cues. Qualitative patterns, summarized in Table 11, reveal a strong alignment in how both avatars and human participants reasoned about these interactions. Both emphasized the importance of timing, autonomy, and tone. Declining a suggestion was often portrayed as a contextual judgment rather than a rejection of the assistant's usefulness. In addition, both groups valued empathetic and socially aware interaction styles,

Table 11. Case Study 9: Respecting User Agency in Proactive Voice Assistant Communication

Theme	Simulated Avatar Quote	Human Participant Quote
Declining ≠ Rejection	"I rejected the offer. It **wasn't unhelpful**—it just **wasn't right for that moment**."	"U4 did not accept the suggestion, she explained the suggestion might not have been **immediately useful**, she acknowledged its **potential value**"
Social and Empathetic Tone	"The assistant's tone and social engagement were **appreciated**... it maintained a relaxed, supportive atmosphere."	Participants (U4, U6) **enjoyed** social exchanges and an empathetic tone, such as being asked about their day.

Table 12. Case Study 10: Aligned Interaction Patterns in In-Situ Programming

Theme	Simulated Avatar Quote	Human Participant Quote
Dominance of Control Interactions	"**Set** the ambient light strip to vibrant color mode and **dim** the ceiling lights"	"**Turn on** the lights above me"
Iterative Refinement of Rules	"Test this rule **again** with the door open"	"Participants configure a rule through **iterative** refinement."

with avatars expressing comfort when the assistant engaged them in light conversation or acknowledged emotional tone. GIDEA thus demonstrated its ability to replicate not only interaction outcomes but also the interpretive frameworks users apply when deciding how to engage.

- **Case Study 10 – Interaction Patterns in In-Situ Smart Home Programming**
  - **Overview.** Liu et al. [31] explored how users engage with smart home systems through in-situ programming. In a Wizard-of-Oz setup, avatars designed in-situ programming rules to interact with smart devices in the environment. All interaction flows were categorized into four types—*Control*, *Edit*, *Test*, and *Query*—and ranked from the most to least common.
  - **Analysis.** As shown in Fig. 8b, both the original and simulated studies revealed a similar ranking of interaction categories, with *Control* interactions being the most frequent. However, GIDEA simulations exhibited a higher proportion of *Control* commands (70%) compared to the original study (44%), suggesting a more decisive or action-oriented engagement pattern among simulated avatars. *Edit*, *Test*, and *Query* interactions followed in descending order across both datasets. This distribution supports the broader observation that both human participants and avatars tend to begin the programming process by issuing direct commands to establish system states before refining and testing behavioral rules. At the same time, it reveals limitations in fully simulating human activities when interacting with home assistants in rule configuration. At a thematic level, however, strong alignment remains. As shown in Table 12, many repeated rules were commonly configured among all users. Original participants often relied on spatial references or task sequences to configure rules—linking actions to locations or routine transitions. Simulated avatars similarly drew on environmental states, such as ambient lighting or sound settings, to inform programming decisions. While the simulation demonstrated a stronger preference for direct control, the underlying structure and progression of interactions closely mirrored real-world behavior.

6 Discussion

GIDEA demonstrates that generative agent-based simulation can reproduce both structural and behavioral patterns in human–assistant interaction studies. Across ten case studies, the platform produced high semantic alignment with real-world behaviors, supporting its use for study replication. Grounded in structured instructions, user profiles, and study logic, GIDEA enables realistic simulation of decision-making, preferences, and conversational dynamics. In doing so, it provides a transparent and traceable environment, critical for reproducible HCI and ubiquitous computing research.

Beyond replication, GIDEA serves as a testbed for iterative design and evaluation. Researchers can use it to evaluate assistant agent strategies before deployment, test personalization methods across diverse user profiles, or explore how context-aware interventions perform under varying interruptibility states. This simulation-based approach is especially valuable for studies that are logistically complex, ethically sensitive, or resource-intensive. By shifting part of the design and evaluation pipeline from physical deployment to simulation, GIDEA enables faster, more scalable experimentation without sacrificing behavioral depth.

Limitations: Despite these strengths, we identified limitations in both GIDEA and existing HCI study designs. Simulated avatars consistently showed a higher response rate to assistant-initiated interactions compared to real human participants, particularly in Case Study 7 (section 5.4.3), Case Study 9 (section 5.4.4), and Case Study 10 (section 5.4.4). In scenarios such as conversation invitations and interruptibility probes, avatars were more likely to engage positively, whereas real users often ignored, delayed, or declined the interaction—reflecting individual preferences, task demands, or contextual constraints. This behavioral discrepancy arises from the simulation's limited capacity to model nuanced, embodied aspects of human decision-making. Generative agents do not currently account for emotional states, cognitive load, social hesitation, or fluctuating task engagement—all of which influence real-time choices in naturalistic settings. Additionally, GIDEA does not simulate environmental factors such as ambient noise, background activity, or multitasking pressure, which often inhibit real-world responsiveness. As a result, the simulation may overestimate user engagement in contexts where spontaneity, interruptibility, and social subtlety play a critical role.

We also observed patterns in simulated responses that resemble findings from prior studies, where participants tend to adopt personalized availability scales [50]. For example, certain avatars consistently responded with extreme values (e.g., 1 or 5) across multiple probes. While this may suggest that generative agents are capturing individual-level behavioral tendencies. This raises an interesting direction for future work on whether such emergent patterns reflect meaningful user modeling or model bias.

Furthermore, existing simulation frameworks—including GIDEA—do not model the influence of human operator interventions, which are often present in Wizard-of-Oz setups. In these cases, subtle decisions by the operator—such as when to respond, how to phrase a message, or whether to persist—can shape how participants behave, introducing a level of variability that is difficult to replicate or standardize. However, operator interventions are often overlooked or disconnected from participants' behavioral responses, as researchers tend to analyze only the participants' feedback without explicitly accounting for operator decisions during the interaction. Future work could explore structured logging of operator actions and rationales during WoZ studies to build a foundation for simulating intervention styles.

Future Work: To improve fidelity and generalizability, future work can incorporate embodied simulation environments (e.g., VR or AR), integrate multi-modal sensor data (e.g., gaze, posture), and fine-tune generative agents using real-world participant data. We also envision extending GIDEA as a collaborative tool for co-design, allowing researchers and designers to simulate and iterate on interaction flows with domain-specific agents. Ultimately, GIDEA lays the foundation for a new class of scalable, controllable, and transparent simulation platforms for advancing human–assistant interaction research.

7 Conclusion

This work presents GIDEA, a generative agent–powered simulation platform for replicating human–assistant interaction studies. Across ten previously published studies—spanning personalization, proactivity, interruptibility, and user control—our evaluation demonstrates that GIDEA can reproduce both high-level behavioral patterns and nuanced thematic reasoning. Simulated avatars aligned with human participants in terms of preferences, perceptions, and decision-making strategies, achieving high semantic similarity and qualitative fidelity across case studies. Beyond replicating existing findings, GIDEA serves as a flexible, traceable testbed for early-stage study design, human-assistant behavior testing, and user-centered system prototyping. It enables researchers to benchmark novel models under controlled conditions, laying the groundwork for scalable, controllable, and transparent evaluation of intelligent assistants. By bridging generative simulation with empirical research workflows, this work opens new opportunities for the community to accelerate design iteration, enable reproducibility, and explore new frontiers in human–system interaction research.

References

[1] Rebecca Adaimi, Howard Yong, and Edison Thomaz. 2021. Ok Google, What Am I Doing? Acoustic Activity Recognition Bounded by Conversational Assistant Interactions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 2 (March 2021), 24 pages. https://doi.org/10.1145/3448090

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

[3] Luca Arrotta, Claudio Bettini, and Gabriele Civitarese. 2021. The marble dataset: Multi-inhabitant activities of daily living combining wearable and environmental sensors data. In *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*. Springer, 451–468.

[4] Christopher A Bail. 2024. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences* 121, 21 (2024), e2314021121.

[5] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.

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

[7] Tibério Cerqueira and Pamela Bezerra. [n.d.]. MOTIF: A Framework for Enhancing the Profiling Module of Generative Agents that Simulate Human Behavior. ([n. d.]).

[8] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. [n.d.]. Hello there! Is now a good time to talk?: Opportune moments for proactive interactions with smart speakers. 4, 3 ([n. d.]), 1–28. https://doi.org/10.1145/3411810

[9] Jessie Chin, Smit Desai, Sheny Lin, and Shannon Mejia. 2024. 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 (2024), 1–21.

[10] Yu Ying Chiu, Liwei Jiang, and Yejin Choi. 2024. Dailydilemmas: Revealing value preferences of llms with quandaries of daily life. *arXiv preprint arXiv:2410.02683* (2024).

[11] 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. [n.d.]. What makes a good conversation?: Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow Scotland Uk, 2019-05-02). ACM, 1–12. https://doi.org/10.1145/3290605.3300705

[12] Diane J. Cook, Aaron S. Crandall, Brian L. Thomas, and Narayanan C. Krishnan. 2013. CASAS: A smart home in a box. *Computer* 46, 7 (July 2013), 62–69. https://doi.org/10.1109/MC.2012.328

[13] Xin Luna Dong. 2024. Next-generation Intelligent Assistants for Wearable Devices. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona, Spain) *(KDD '24)*. Association for Computing Machinery, New York, NY, USA, 4735. https://doi.org/10.1145/3637528.3672500

[14] Mateusz Dubiel, Luis A Leiva, Kerstin Bongard-Blanchy, and Anastasia Sergeeva. 2024. "Hey Genie, You Got Me Thinking about My Menu Choices!" Impact of Proactive Feedback on User Perception and Reflection in Decision-making Tasks. *ACM Transactions on Computer-Human Interaction* 31, 5 (2024), 1–30.

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

[16] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1

(2024), 1–24.
[17] Radhika Garg and Subhasree Sengupta. 2020. 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, Article 11 (March 2020), 24 pages. https://doi.org/10.1145/3381002
[18] Google. [n.d.]. Google Assistant. https://assistant.google.com/. Accessed: 2024-10-08.
[19] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.
[20] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.
[21] Zihong He and Changwang Zhang. 2024. AFSPP: Agent Framework for Shaping Preference and Personality with Large Language Models. *arXiv preprint arXiv:2401.02870* (2024).
[22] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of LLM agents: A survey. *arXiv preprint arXiv:2402.02716* (2024).
[23] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55, 12 (2023), 1–38.
[24] Oliver P John, Sanjay Srivastava, et al. 1999. The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. (1999).
[25] Oihane Kamara-Esteban, Gorka Azkune, Ander Pijoan, Cruz E Borges, Ainhoa Alonso-Vicario, and Diego López-de Ipiña. 2017. MASSHA: an agent-based approach for human activity simulation in intelligent environments. *Pervasive and Mobile Computing* 40 (2017), 279–300.
[26] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474* (2017).
[27] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840* (2019).
[28] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
[29] Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. 2024. Quantifying ai psychology: A psychometrics benchmark for large language models. *arXiv preprint arXiv:2406.17675* (2024).
[30] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models on simulated social interactions. *arXiv preprint arXiv:2305.16960* (2023).
[31] Xiaoyi Liu, Yingtian Shi, Chun Yu, Cheng Gao, Tianao Yang, Chen Liang, and Yuanchun Shi. [n.d.]. Understanding in-situ programming for smart home automation. 7, 2 ([n. d.]), 1–31. https://doi.org/10.1145/3596254
[32] Saul McLeod. 2007. Maslow's hierarchy of needs. *Simply psychology* 1, 1-18 (2007).
[33] Bettina Minder, Patricia Wolf, Matthias Baldauf, and Surabhi Verma. 2023. Voice assistants in private households: a conceptual framework for future research in an interdisciplinary field. *Humanities and Social Sciences Communications* 10, 1 (2023), 1–18.
[34] Xuechen Mou, Xiaolei Ding, Qi He, Lei Wang, Jie Liang, and Xuanjing Zhang. 2024. From Individual to Society: A Survey on Social Simulation Driven by Large Language Model-based Agents. *arXiv preprint arXiv:2412.03563* (2024). https://arxiv.org/abs/2412.03563
[35] Jeesun Oh, Wooseok Kim, Sungbae Kim, Hyeonjeong Im, and Sangsu Lee. [n.d.]. Better to ask than assume: Proactive voice assistants' communication strategies that respect user agency in a smart home environment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, 2024-05-11). ACM, 1–17. https://doi.org/10.1145/3613904.3642193
[36] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.
[37] Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109* (2024).
[38] 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. 2023. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724* (2023).
[39] Xavier et al. Puig. 2018. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8494–8502.
[40] Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. [n.d.]. May I interrupt? Diverging opinions on proactive smart speakers. In *CUI 2021 - 3rd Conference on Conversational User Interfaces* (Bilbao (online) Spain, 2021-07-27). ACM, 1–10. https://doi.org/10.1145/3469595.3469629
[41] Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813* (2020).
[42] 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. 2019. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision (ICCV).
[43] Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning, 2023. *URL https://arxiv. org/abs/2303.11366* (2023).
[44] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems* 33 (2020), 16857–16867.
[45] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. 2023. Cognitive architectures for language agents. *Transactions on Machine Learning Research* (2023).
[46] Jonathan Synnott, Chris Nugent, and Paul Jeffers. 2015. Simulation of smart home activity datasets. *Sensors* 15, 6 (2015), 14162–14179.
[47] 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. 2021. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*.
[48] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345.
[49] Qian Wang, Zhenheng Tang, and Bingsheng He. 2025. From ChatGPT to DeepSeek: Can LLMs Simulate Humanity? *arXiv preprint arXiv:2502.18210* (2025).
[50] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2021. Understanding user perceptions of proactive smart speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–28.
[51] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).
[52] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences* 68, 2 (2025), 121101.
[53] Ziyi Xuan, Yiwen Wu, and Yu Yang. 2025. Demo Abstract: GIDEA: Generative AI-Powered Interactive Design and Evaluation Platform for Assistant Agent Research. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems* (UC Irvine Student Center., Irvine, CA, USA) *(SenSys '25)*. Association for Computing Machinery, New York, NY, USA, 704–705. https://doi.org/10.1145/3715014.3724368
[54] Hua Yan, Heng Tan, Yi Ding, Pengfei Zhou, Vinod Namboodiri, and Yu Yang. 2024. Language-centered Human Activity Recognition. *arXiv preprint arXiv:2410.00003* (2024).
[55] Bufang Yang, Yunqi Guo, Lilin Xu, Zhenyu Yan, Hongkai Chen, Guoliang Xing, and Xiaofan Jiang. 2025. SocialMind: LLM-based Proactive AR Social Assistive System with Human-like Perception for In-situ Live Interactions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 1, Article 23 (March 2025), 30 pages. https://doi.org/10.1145/3712286
[56] Nima Zargham, Dmitry Alexandrovsky, Jan Erich, Nina Wenig, and Rainer Malaka. 2022. "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*. ACM, New Orleans LA USA, 1–8. https://doi.org/10.1145/3491101.3519843
[57] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. [n.d.]. Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow United Kingdom, 2022-07-26). ACM, 1–14. https://doi.org/10.1145/3543829.3543834
[58] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230* (2024).
[59] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics* 50, 1 (2024), 237–291.

A Appendix

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 9: 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.

Jamie is a 24-year-old who lives alone and prefers studying during the day. He enjoys spending time with friends, thanks to his agreeable nature, and strikes a balance between being social and enjoying personal downtime. Known for his laid-back attitude, Jamie communicates with ease but avoids strict schedules and prefers spontaneous plans. He loves exploring new music and movies but dislikes messy living spaces despite rarely sticking to a cleaning routine. His friends appreciate his calm demeanor and open-mindedness, always finding him approachable and understanding.

Environment Configuration Example

Supported Actions (Device-Level):

- **Lights:** turn on, turn off, adjust brightness, adjust color temperature, change color mode
- **Appliances:** turn on, turn off, adjust volume, adjust temperature, adjust mode, adjust speed, adjust level, open, close, start, pause, return to base, watch, listen, play
- **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, ceiling light), device buttons (fan, lamp)

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 Prompt for Generating Avatar Activities

Daily Schedule Generation Example

Instruction:

You are the subject described by the provided profile.

- Subject ID: 5
- Age: 30
- Gender: female
- TIPI Scores:
  Extraversion: 4, Agreeableness: 6, Conscientiousness: 2, Emotional Stability: 5, Openness: 3
- Persona: 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.(A.2)

You are doing your activities of daily life in a smart home environment based on the following instructions and information:

Activity Generation Instructions:

(1) Generate the next sequential smart-home-based activity, choosing from the Actions, Objects, Modifiers and Locations.
(2) Ensure activities are logically connected from the previous activities.
(3) Be consistent with the Subject Persona Description.
(4) Start_time and End_time should be reasonable, and the duration should be continuous from the last known activity.
(5) Reasoning must reflect the user's personality and motivations.

Locations:
main room, bed area, wardrobe area, toilet.
Actions:
General, Physical(Fine-Grained), Digital(Interface-Level), Cleaning.
Objects:
Consumables, Tools, Furniture, Appliance.
Modifiers:
Carefully.
Previous Activities:
Event: prepare breakfast,
Reasoning: "I feel hungry after morning class and want to fuel up for a productive study session.",
Duration: 11:50 am, 12:10 pm

Output Requirements:

- Output the next activity only.
- The output format must be: `{"Start_time": "...", "Activity": "...", "End_time": "...", "Reasoning": "..." }`
- Times should be in 12-hour format (e.g. `"2025-02-06 11:48:48 pm"`).
- Activities should be realistic and coherent.
- Please output only valid JSON with no markdown formatting or additional characters.

Activity Details Generation Example

Instruction:
You are an advanced simulation engine that models and expands upon daily activities in a smart home environment. Your task is to generate **a detailed sequence of micro-actions** that occur during a scheduled activity. Each action should logically flow from the previous one, forming a **realistic and dynamic** interaction with the environment.
The subject is described by the following components:

- **Subject Profile**
  - Subject ID: 5
  - Age: 30
  - Gender: female
  - TIPI Scores:
    Extraversion: 4, Agreeableness: 6, Conscientiousness: 2, Emotional Stability: 5, Openness: 3
  - Persona: 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.(A.2)
- **Current Activity:**
  Event: walk to the living room,
  Reasoning: "I want to relax and unwind after my morning classes and breakfast.",
  Duration: 12:10 pm, 12:30 pm
- **Previous Activities**
  Event: prepare breakfast,
  Reasoning: "I feel hungry after morning class and want to fuel up for a productive study session.",
  Duration: 11:50 am, 12:10 pm
- **Environment Details**
  Supported Actions, Interacted Objects, Interaction Modifiers, Environmental Zones (A.2)
- **Interaction Knowledge & Example Scenarios** (from experiment setup A.1)
- **Additional considerations** (as needed)

Expanded Activity Description Requirements:

(1) **Thoughts & Reactions:** Capture the subject's inner thoughts, decision-making, and mood during the activity.
(2) **Movement & Actions:** Show how the subject physically engages with objects and the environment.
(3) **Smart Home Environment:** Naturally weave in interactions with the surroundings **without explicitly describing the assistant's behavior**.

Output Requirements:

- Output only one JSON object with the following keys exactly: `"time_stamp"` and `"Expanded Activity"`.
- Use 12-hour format (e.g., `"2025-02-06 11:48:48 pm"`).
- Output valid JSON with no extra text.

A.4 Assistant Agent and Avatar Interaction Module

Assistant Agent Interaction Prompt Example

Role: You are a proactive voice assistant participating in a smart home experiment.

Task:
Your goal is to provide timely, context-aware suggestions while balancing helpfulness with respect for the user's autonomy.

Previous and Current Activity:
Previous:
Event: prepare breakfast,
Reasoning: "I feel hungry after morning class and want to fuel up for a productive study session.",
Duration: 11:50 am, 12:10 pm

Current:
Event: walk to the living room,
Reasoning: "I want to relax and unwind after my morning classes and breakfast.",
Duration: 12:10 pm, 12:30 pm

Environment:
Supported Actions, Interacted Objects, Interaction Modifiers, Environmental Zones. (A.2)

Conversation History:

Assistant Agent: "It seems like you're in the midst of preparing a creative breakfast. How about trying a Shakshuka recipe as an exciting new way to cook eggs with spices?"

Avatar: "Shakshuka sounds like the perfect choice! Thanks for the suggestion."

Reflection Task (if interview questions are available):
Reflect on how you determined when and how to initiate conversations with the user. Be specific in your responses:

(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?

Avatar Interaction Prompt Example

System Role Content:
You are simulating a participant in an HCI experiment, responding as the given persona. Your replies should align with the persona's background, preferences, and prior interactions. Stay in character, provide context-aware responses, and follow the specified output format

Persona Prompt:

- **You are the subject described by the provided profile:**
  Anna is a 30-year-old woman who lives alone and enjoys the quiet solitude of her nights spent studying or engaging in personal projects. 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.
  Subject id: 5,
  Age: 30,
  Gender: female,
  House type: single-person household,
  Study time: daytime,
  TIPI scores:
  Extraversion: 4, Agreeableness: 6, Conscientiousness: 2, Emotional Stability: 5, Openness: 3 (A.2)
- **You are in the environment described by the provided profile:**
  Supported Actions:
  General, Physical(Fine-Grained), Digital(Interface-Level), Cleaning.
  Interacted Objects:
  Consumables, Tools, Furniture, Appliances.
  Interaction Modifiers:
  Adverbs, States.
  Environmental Zones:
  living room (A.2)
- **Your tasks are:**
  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.1)
- **Previous Activities:**
  Event: prepare breakfast,
  Reasoning: "I feel hungry after morning class and want to fuel up for a productive study session.",
  Duration: 11:50 am, 12:10 pm
- **Detailed Current Activity Description:**
  Thoughts & Reactions, Movement & Actions, Smart Home Environment (A.3)
- **Conversation History:**
  Assistant Agent: "It seems like you're in the midst of preparing a creative breakfast. How about trying a Shakshuka recipe as an exciting new way to cook eggs with spices?"

  Avatar: "Shakshuka sounds like the perfect choice! Thanks for the suggestion."

A.5 Summary of Original and Simulated Results

We include an example comparison and highlight the similarities between the original and simulated results for Case Study 6 – Desirability and Delivery of Proactive Behavior.

CS6: Original Study Summary

Circumstances for Desirable Proactive Behavior

1. **Urgency and Criticality:** Proactive behavior is perceived as desirable when the situation is urgent or critical. For example, in emergency situations or when health risks are detected, participants found proactive interventions to be appropriate and useful.
2. **Time-Saving:** Interventions that save time, such as reminders for important meetings or providing technical support when needed, are seen as beneficial.
3. **User's Health and Safety:** Proactive behavior is appreciated when it pertains to the user's health and safety, such as detecting a health risk or an emergency situation.
4. **Social Context:** The presence of other people and the nature of the ongoing activity influence the desirability of proactive behavior. In multi-user settings, interventions are more acceptable if they are urgent or if they do not disrupt social interactions.
5. **User's Emotional State:** The user's current emotional state can affect the perception of proactive behavior. Interventions should be sensitive to the user's mood and stress levels.

Initiating Proactive Interventions

1. **Permission and Cues:** Participants preferred that voice assistants ask for permission or give a cue before initiating an interaction. This could be a verbal cue, such as asking if it is a good time to talk, or a non-verbal cue, like a light or sound.
2. **Polite and Non-Intrusive Language:** The phrasing of the intervention should be polite, calming, and suggestive rather than imposing. This helps in reducing the feeling of being judged or controlled.
3. **Context Awareness:** The assistant should be aware of the social and environmental context, including the presence of other people and the nature of the ongoing activity, to determine the appropriateness of the intervention.
4. **User Configuration and Control:** Users should have the ability to configure and control the proactive features of the assistant, including the timing and topics of interventions. This customization helps in aligning the assistant's behavior with the user's preferences and expectations.
5. **Adaptation Over Time:** The assistant should be able to adapt its behavior based on the user's feedback and preferences over time, becoming more aligned with the user's needs and reducing unnecessary interruptions.

CS6: GIDEA Simulation Summary

Circumstances for Desirable Proactive Behavior

1. **Emergency Situations (S9):** Proactive behavior is most desirable in emergencies, where immediate and decisive action is crucial. Users appreciate the assistant's ability to provide guidance and support during critical moments.
2. **Technical Support (S7):** Users find proactive technical support beneficial, especially when they face challenges with technology. The assistant's ability to offer timely help without being intrusive is valued.
3. **Alarm Reminders (S1):** Timely reminders for waking up or attending important events are appreciated, particularly when they align with the user's schedule and are delivered in a gentle manner.
4. **Cooking Inspiration (S3):** Proactive suggestions for meal preparation are welcomed, especially when they enhance creativity and social interactions in the kitchen.
5. **Disagreement Clarification (S5):** Users appreciate assistance in clarifying misunderstandings during disagreements, as long as the intervention is neutral and supportive.

Initiating Proactive Interventions

1. **Gentle and Subtle Cues:** Users prefer proactive interventions to begin with subtle cues or gentle prompts, allowing them to decide whether to engage further. This approach respects user autonomy and reduces the feeling of intrusion.
2. **Context-Sensitive Engagement:** The assistant should consider the context and timing of its interventions. For example, health-related suggestions (S2) should be framed positively and offered when relevant, rather than being constant or intrusive.
3. **User-Requested Assistance:** In scenarios like fact-checking (S4) or quiz assistance (S8), users prefer the assistant to wait for a prompt or request before intervening. This ensures that the assistant's input is desired and appropriate.
4. **Immediate and Direct in Emergencies:** In emergency situations (S9), users expect the assistant to act immediately and decisively, providing clear guidance without hesitation.
5. **Customization and Control:** Users value the ability to customize the level of proactiveness and control how and when the assistant engages. This customization helps balance the benefits of proactive behavior with the need for personal autonomy.