Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 470–480
https://doi.org/10.54941/ahfe1006859

Data-Grounded Empathy: Simulating "The Untouchable" to Mitigate Representational Bias in User Research

Xiaoman Lin, Yufei Wang, Leran Zhou, Anzhe Huang, and Yunmao Gao

Xi'an Jiaotong University, Xi'an, China

ABSTRACT

Traditional user research in high-pressure service contexts is often constrained by logistical challenges and the pervasive influence of social desirability bias (SDB), which compromises data authenticity. This paper presents a reproducible workflow for developing and validating a high-fidelity AI interview agent designed to address these challenges. Built on a Retrieval-Augmented Generation (RAG) architecture, the agent is grounded in a multi-source knowledge base compiled from in-depth interviews, online community discussions, and multimedia content from service-industry workers. We describe the end-to-end process, from data collection and preprocessing to agent implementation and prompt engineering. The agent's performance was assessed through a two-part validation study: an expert heuristic evaluation and a comparative Turing test involving 22 participants. The results show that the agent produced interview data that were perceptually indistinguishable from human-generated responses and were rated by participants as significantly more consistent and coherent. This work contributes a transparent and adaptable methodology for Human–Computer Interaction (HCI) and design research, offering a scalable tool to gather authentic user insights while mitigating known biases. The findings point to a new paradigm for human–AI collaboration in user research, particularly for accessing hard-to-reach populations.

Keywords: AI agents, User research, Conversational AI, Social desirability bias, Retrieval-augmented generation (RAG), Reproducibility, Human-computer interaction

INTRODUCTION

The Challenge of Authentic Data in User Research

Qualitative user research is central to human-centered design because it reveals users' behaviours, needs, and motivations (McCue et al., 2023). However, obtaining authentic data remains difficult, especially in high-pressure service industries (Lv et al., 2024).

Time, cost, and geographic barriers often limit participant recruitment and research scope. More importantly, social desirability bias (SDB)—the tendency to respond in socially acceptable ways—can distort findings (Tourangeau & Yan, 2007). This bias is especially strong when discussing

sensitive workplace issues like conflict, dissatisfaction, or stress (Boring & Delfgaauw, 2024).

The presence of a human interviewer may intensify this effect as participants alter responses to avoid judgment, leading to incomplete or idealized accounts of user experience (Schumann & Lück, n.d.).

AI Agents as a Methodological Intervention

Advances in conversational AI provide new tools for HCI research (Lv et al., 2025). Properly designed AI interview agents address several limitations of traditional methods (Budig et al., 2025). They enable scalable and accessible data collection and offer a neutral, nonjudgmental environment that reduces social pressure and encourages candid disclosure (Chen & Li, 2017).

Research Contribution

This paper proposes a new approach to using AI for authentic and scalable user research. The contributions are threefold:

1. **A Reproducible Workflow:** An end-to-end process for designing and validating an AI interview agent, grounded in real-world data via a RAG architecture, improving transparency and reproducibility (Alaofi et al., 2025).
2. **Empirical Validation:** Evidence from mixed-method validation showing that the agent achieves high fidelity and generates responses perceptually indistinguishable from human data while demonstrating superior coherence.
3. **Insights on AI-Mediated Authenticity:** Insights into how AI agents promote candid responses, emphasizing that their value lies not in imitation but in consistent, bias-resistant interaction.

RELATED WORK

Conversational Agents in Research and Design

Conversational agents (CAs) have evolved from simple command systems into collaborators in research and design workflows (Budig et al., 2025). In HCI, an agent's persona—its tone, personality, and embodiment—strongly influences user trust and engagement (Janson, 2023). Recent studies show that AI-based surveys and interviews can increase engagement and elicit richer responses compared to static questionnaires, supporting their use for qualitative research (Jacobsen et al., 2025).

Social Desirability Bias and Interview Modality

SDB manifests as either unintentional self-deception or deliberate impression management (Seitl et al., 2024). Interview format significantly affects bias levels: the presence of a human interviewer, especially face-to-face, can heighten awareness of social norms and suppress negative expression (Teh et al., 2023).

This raises a key question for AI interviewer design: What level of human-likeness is optimal? Studies suggest that highly anthropomorphic agents may trigger social responses reinforcing SDB (Teh et al., 2023), while neutral and nonjudgmental agents foster more honest disclosure, particularly on sensitive topics (Jo, 2024). Hence, research-oriented agents should balance human-like engagement with neutrality.

Synthetic Users and Agent-Based Simulation

This work aligns with emerging research on synthetic users and agent-based simulation (Park, 2023). Historically, user simulation has been employed for the automated, quantitative evaluation of systems like conversational recommender systems (CRSs) (Afzali et al., 2023). Advances in large language models (LLMs) now allow generative agents to approximate human reasoning and subjective judgment (Park, 2023).

Unlike prior studies focused on system testing, this study uses an AI agent to generate qualitative interview data, aligning with the Synthetic User concept—a data-grounded, interactive persona for hypothesis testing and empathetic understanding.

Grounding Generation With RAG

Foundational LLMs often hallucinate and lack empirical grounding. The Retrieval-Augmented Generation (RAG) framework mitigates this by conditioning outputs on verified data (Fan et al., 2024). Here, the RAG structure ensures that the agent's responses reflect real service worker experiences rather than generic LLM output, improving both accuracy and representational authenticity (Yang et al., 2025).

METHODOLOGY: A RAG-POWERED AI AGENT WORKFLOW

Our research proposes a systematic, four-stage workflow for the development and validation of an AI interview agent. This process is designed to be transparent and reproducible. Figure 1 provides a high-level overview of the workflow.

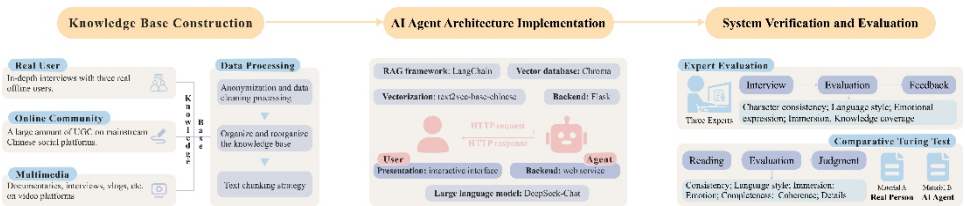

Figure 1: The workflow for developing the RAG-based AI interview agent, from data collection to system deployment and validation.

Stage 1: Multi-Source Knowledge Base Construction

The fidelity of a RAG-based agent depends on the quality and diversity of its knowledge base. To build a comprehensive representation of service industry workers, we adopted a multi-source data collection strategy.

- **Data Sources**
  - **In-depth Interviews:** Three semi-structured interviews with service workers captured detailed and emotional first-hand accounts of their work experiences.
  - **Online Community Data:** User-generated posts from Chinese social media platforms (Zhihu, Xiaohongshu) provided anonymous and candid group-level perspectives, including common challenges and shared experiences.
  - **Multimedia Content:** Transcripts from documentaries and vlogs (Bilibili, Douyin) offered contextual cues and emotional expressions often missing in text-only data.

- **Preprocessing Pipeline**
  - **Anonymization and Cleaning:** All data were anonymized to remove personally identifiable information (PII) and cleaned to eliminate duplicates and irrelevant content.
  - **Thematic Annotation and Chunking:** The processed data were manually labeled into 62 fine-grained themes (e.g., customer conflict, scheduling issues, colleague relationships) and segmented into coherent chunks of 200–300 characters for effective retrieval in the RAG system.

Stage 2: AI Agent Architecture and Implementation

The agent was built using a modular, three-tier architecture to ensure robustness and scalability. Figure 2 illustrates the system's data flow.

- **System Architecture:** The system comprises three layers: a User Layer (researcher-facing interface), an Application Layer (Flask-based backend managing requests and dialogue states), and a Core Agent Layer (RAG module handling retrieval and generation).
- **Technology Stack:** We employed open-source tools, including LangChain as the RAG framework, text2vec-base-chinese for text embeddings, Chroma as the vector database, DeepSeek-Chat as the core LLM, and Flask for backend web services.
- **RAG Mechanism:** Upon receiving a user query, the input is converted into a 768-dimensional vector. A cosine similarity search is conducted in the Chroma database to retrieve the top-K relevant knowledge chunks. Empirical testing showed $K = 20$ achieved the best balance between contextual sufficiency and noise. These chunks are concatenated into a context block and passed to the LLM for generation.

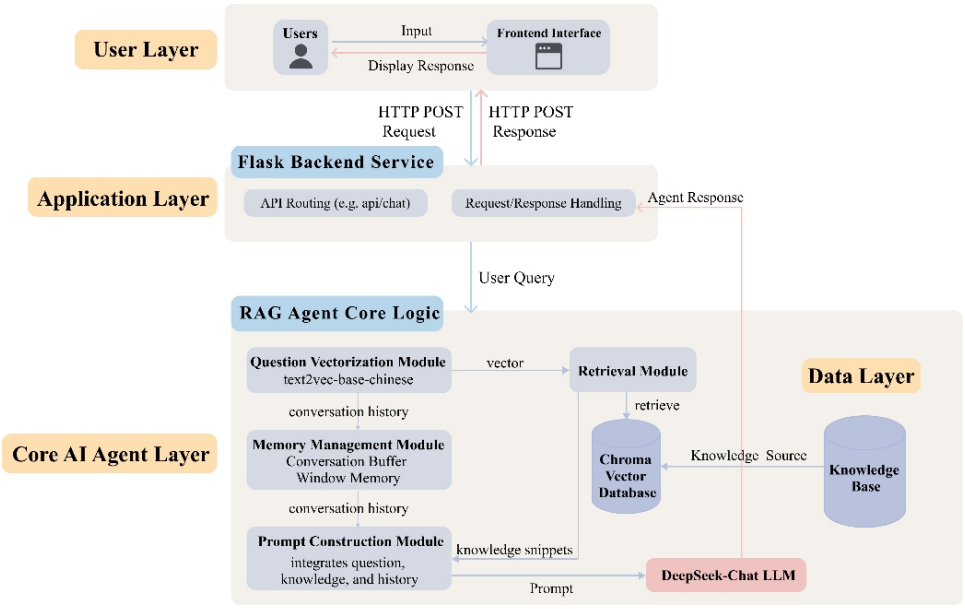

Figure 2: System architecture and data flow of the AI agent, showing the interaction between the user interface, backend application, and the core RAG module.

Stage 3: Persona Crafting via Prompt Engineering

The prompt is the key element shaping the agent's personality and behaviour. Its design follows three principles: persona consistency, linguistic and emotional authenticity, and response moderation. The prompt includes three components:

1. **Role Definition:** Defines the agent's identity, including its name (Xiaomei), age (22), and work experience (hot pot restaurant server).
2. **Language & Emotional Style:** Specifies tone (warm, approachable), use of colloquial expressions, concise sentence structure (short, concise), and examples for expressing emotions naturally.
3. **Behavioral Rules:** Sets strict output constraints, requiring first-person narration, synthesis of retrieved knowledge instead of repetition, and concise responses to avoid verbosity.

Stage 4: Maintaining Coherence With Multi-Turn Memory

To maintain conversational coherence, we implemented a multi-turn memory mechanism that retains a sliding window of the last 25 dialogue turns (user queries and agent responses). This history is appended to each new prompt as contextual input, enabling the agent to reference prior interactions and sustain a logical flow. The approach preserves context effectively while preventing token overload from an ever-expanding conversation history.

VALIDATION STUDY

Study Design and Rationale

To assess the AI agent's performance, we designed a mixed-method, two-part validation study. The goal was to evaluate both its fidelity (how believably it simulates the target persona) and its effectiveness (the quality and utility

of its responses for user research purposes). This dual approach integrates qualitative expert evaluation with quantitative user perception data for a comprehensive assessment.

Study 1: Expert Heuristic Evaluation

- **Participants:** Three experts with master's degrees and professional backgrounds in HCI, UX research, and conversational AI/NLP participated. All had substantial experience conducting qualitative interviews.
- **Procedure:** Each expert conducted a full semi-structured interview with the AI agent "Xiaomei" following a predefined guide covering topics such as daily work, job satisfaction, and interpersonal relationships. After the interview, they completed a heuristic evaluation questionnaire.
- **Measures:** The questionnaire assessed the agent's performance on five heuristic dimensions using 5-point Likert scales: Persona Consistency, Language Style, Immersion, Knowledge Coverage, and Emotional Expression.

Study 2: Comparative Turing Test

This study adapted the classic Turing Test into a comparative evaluation (Rathi et al., 2024). Instead of asking participants to identify human versus machine responses, they were asked to rate the quality of transcripts without knowing their source. This approach offers a more nuanced measure of perceived authenticity and quality, aligning with contemporary interpretations of the test's relevance in the age of advanced LLMs (Jo, 2024).

- **Participants:** Twenty-two participants from diverse backgrounds were recruited to ensure a broad range of perceptual judgments.
- **Materials:** Two interview transcripts of equivalent length and topic coverage were prepared:
  - **Interview A:** A verbatim transcript from a real service worker interview.
  - **Interview B:** Transcript generated from an AI agent interview.
- **Procedure:** Participants were randomly assigned to read either Interview A or B first. After reading each transcript, they completed an evaluation questionnaire and then made a forced-choice judgment on whether each transcript was produced by a human or AI.
- **Measures:** The questionnaire used 5-point Likert scales to rate each transcript on seven dimensions: Persona Consistency, Language Style, Immersion, Content Completeness, Emotional Expression, Content Coherence, and Detail & Specificity.

RESULTS

Expert Evaluation: High Fidelity and Groundedness

The AI agent received consistently high scores from the expert evaluators, confirming its strong fidelity. Persona Consistency ($M = 4.83$, $SD = 0.29$)

and Language Style (M = 4.67, SD = 0.29) were rated highest, indicating that the prompt design effectively produced a coherent and believable character. Scores for Knowledge Coverage (M = 4.50, SD = 0.50) and Emotional Expression (M = 4.34, SD = 0.29) were also strong.

Qualitative feedback supported these results. Experts noted that the agent maintained a stable persona and grounded its responses in realistic scenarios. The main limitations were inherent to the RAG framework: its knowledge was constrained by the database, and some emotional expressions, though appropriate, occasionally appeared formulaic.

Comparative Turing Test: Indistinguishable and Superiorly Coherent

The comparative Turing test provided strong evidence for the agent's ability to generate human-like text.

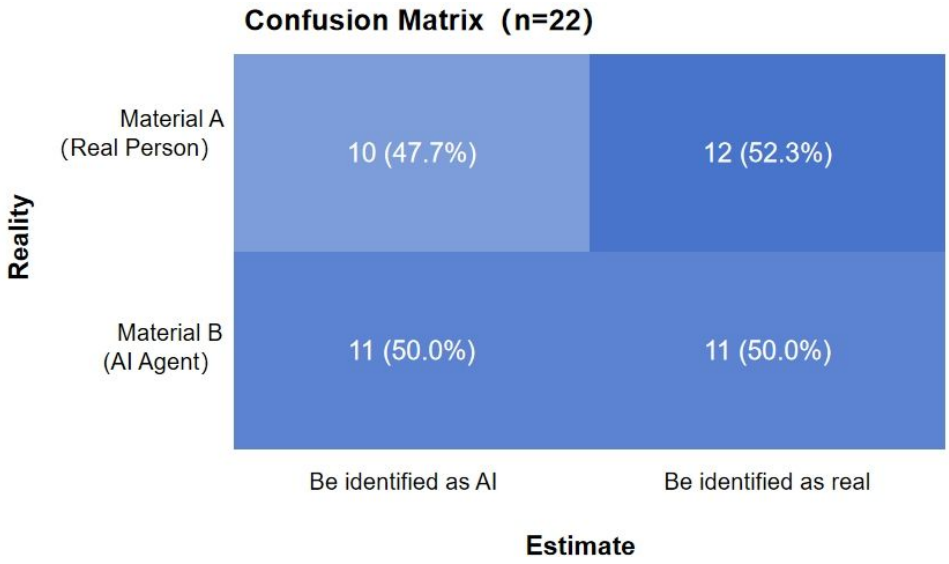

Figure 3: Confusion matrix for the source identification task (n = 22). The results show that participants' judgments were at chance level.

- **Indistinguishability:** Participants were unable to reliably distinguish the AI-generated transcript from the human one. As shown in Figure 3, the AI transcript was identified as human 52.3% of the time, while the human transcript was identified as AI 47.7% of the time. The overall classification accuracy was 50%, which is at chance level.
- **Comparative Quality:** Paired-samples t-tests compared ratings of the AI and human transcripts across seven quality dimensions (Table 1). There were no significant differences in Language Style, Immersion, Content Completeness, or Detail & Specificity (all $p > .05$), suggesting that the AI responses were perceived as equally natural, engaging, and informative as human ones.

Notably, the AI agent scored significantly higher than the human respondent on Persona Consistency ($t(21) = -2.42$, $p = .025$), Emotional

Expression ($t(21) = -2.19$, $p = .041$), and Content Coherence ($t(21) = -3.81$, $p = .001$). These results indicate that, in structured interviews, the agent's responses were perceived as more stable, coherent, and emotionally appropriate.

Table 1: Comparative analysis of participant ratings for human (material A) and AI (material B) interview transcripts (p <.05*, p <.01**).

Evaluation Dimension	Material A: M (sd)	Material B: M (sd)	t(21)	p-Value	Cohen's D
Persona Consistency	4.13 (0.77)	4.55 (0.61)	–2.42	0.025*	–0.52
Language Style	4.05 (0.62)	4.27 (0.82)	–1.29	0.213	–0.27
Immersion	4.21 (0.60)	4.23 (0.69)	–0.13	0.897	–0.03
Content Completeness	4.45 (0.59)	4.31 (0.64)	1.04	0.310	0.22
Emotional Expression	4.09 (0.61)	4.47 (0.51)	–2.19	0.041*	–0.47
Content Coherence	4.18 (0.60)	4.74 (0.40)	–3.81	0.001**	–0.81
Detail & Specificity	4.36 (0.62)	4.18 (0.74)	1.10	0.282	0.23

DISCUSSION

Implications for HCI and Design Research

The workflow proposed in this study offers a practical and scalable method for creating interviewable digital personas or interactive synthetic users. Researchers can interact with these agents on demand, enabling rapid exploration and hypothesis testing without the logistical constraints of human recruitment.

This aligns with broader trends in design research that use AI to enhance user analysis, concept generation, and evaluation (Takaffoli et al., 2024). Integrating synthetic agents in early design stages allows researchers to test assumptions and refine hypotheses before involving human participants, improving both scalability and methodological rigor.

Rethinking Authenticity: The Value of "Super-Human" Consistency

A key finding is that the AI agent was not only comparable to humans but was perceived as more consistent and coherent. This challenges the aim of replicating human imperfection. Human conversation is often fragmented and influenced by fatigue or emotion, whereas the RAG-based agent draws from a stable knowledge base and follows consistent behavioural rules.

This super-human consistency addresses the anthropomorphism paradox: while human-like qualities aid engagement, excessive realism can reintroduce social pressures that trigger SDB. The agent's computational reliability provides a predictable conversational space that promotes openness and honest disclosure. Its value lies not in mimicking human flaws but in offering an idealized, bias-resistant persona.

A Contribution to Reproducibility in AI-Driven Research

Reproducibility remains a major challenge in AI and HCI, where opaque systems hinder validation. This study contributes by presenting a transparent

workflow covering data collection, preprocessing, architecture design, and prompt engineering. By documenting each stage, it provides a replicable framework that others can adapt. Such openness contrasts with the black-box nature of many AI systems and supports calls for standardization in data-intensive research.

Limitations and Future Work

Despite its promising outcomes, this approach has several limitations that suggest directions for future research.

- **Knowledge Base Dependency:** The agent's knowledge remains static and confined to its pre-constructed dataset. It cannot generate genuinely novel insights or adapt to topics outside its defined knowledge scope. This constraint also introduces potential algorithmic bias, as any bias embedded in the original data can propagate through the model.
- **Absence of Non-Verbal Cues:** The current text-only interaction lacks the nonverbal dimensions—such as tone, prosody, and body language—that are integral to in-person qualitative interviews. Integrating multimodal interaction (e.g., voice or embodied avatars) could enhance emotional realism and participant engagement.
- **Risk of Averaging:** By synthesizing responses from multiple data sources, the agent tends to represent an averaged or composite user. While this supports generalizability, it may obscure outlier or contradictory experiences that often spark design innovation.

Future research should explore dynamic and multimodal knowledge bases and develop no-code interfaces for customizable agents. Applying this workflow to other hard-to-reach populations could further demonstrate its scalability and value.

Ethics and Privacy Considerations

The use of AI to simulate human subjects introduces significant ethical responsibilities, which must be addressed with care and transparency (Agnew et al., 2024).

- **Data Sourcing and Privacy:** A cornerstone of this study's methodology was the ethical management of data. Primary participants gave informed consent; secondary data were drawn only from public sources. All data were anonymized to remove identifiable details while preserving context.
- **Informed Consent:** Public data without explicit consent raise ethical concerns due to contextual sensitivity. In this study, data were aggregated to reflect collective experiences and stripped of identifiers. This aligns with digital ethnography practices but requires continued ethical reflection.
- **Risk of Over-Reliance and De-skilling:** Another ethical concern involves the potential over-reliance on AI agents, which could lead researchers and designers to undervalue direct engagement with real users. Excessive dependence on synthetic participants may erode the empathy and contextual sensitivity that are fundamental to human-centered design.

To mitigate this risk, we position the AI agent as a complementary research tool—ideal for early-stage exploration, hypothesis generation, and scaling qualitative insights—rather than as a replacement for human-centered fieldwork. Maintaining a hybrid approach ensures that technological efficiency does not come at the cost of human connection or interpretive depth.

CONCLUSION

This study addressed challenges in obtaining authentic data in the service industry by developing and validating an AI interview agent based on a multi-source knowledge base and a Retrieval-Augmented Generation (RAG) architecture. The agent produced interview data indistinguishable from human responses and rated higher in consistency and coherence.

The proposed workflow provides a transparent and reproducible framework that enhances efficiency and access to hard-to-reach populations. More broadly, it advances human–AI collaboration in user research, showing that AI's non-human strengths—especially its consistency—can help mitigate social desirability bias. As AI further integrates into design practice, thoughtfully designed agents like this may support scalable and bias-aware user research.

REFERENCES

Afzali, J., Drzewiecki, A. M., Balog, K., & Zhang, S. (2023). UserSimCRS: A User Simulation Toolkit for Evaluating Conversational Recommender Systems. Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, 1160–1163.

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). The Illusion of Artificial Inclusion. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 1–12.

Alaofi, M., Arabzadeh, N., Clarke, C. L. A., & Sanderson, M. (2025). Generative Information Retrieval Evaluation (Vol. 51, pp. 135–159).

Boring, A., & Delfgaauw, J. (2024). Social desirability bias in attitudes towards sexism and DEI policies in the workplace. Journal of Economic Behavior & Organization, 225, 465–482.

Budig, T., Nißen, M., & Kowatsch, T. (2025). Towards the Embodied Conversational Interview Agentic Service ELIAS: Development and Evaluation of a First Prototype. Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, 420–424.

Chen, H.-T., & Li, X. (2017). The contribution of mobile social media to social capital and psychological well-being: Examining the role of communicative use, friending and self-disclosure. Computers in Human Behavior, 75, 958–965.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 6491–6501.

Jacobsen, R. M., Cox, S. R., Griggio, C. F., & van Berkel, N. (2025). Chatbots for Data Collection in Surveys: A Comparison of Four Theory-Based Interview Probes. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1–21.

Janson, A. (2023). How to leverage anthropomorphism for chatbot service interfaces: The interplay of communication style and personification. Computers in Human Behavior, 149, 107954.

Jo, E. (2024). Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 1–21.

Lv, X., Gu, Y., Solomon, O. M., Shen, Y., Ren, Y., & Wei, Y. (2024). Status and influencing factors of nurses' organizational silence in general hospitals in eastern coastal cities of China. BMC Nursing, 23(1), 757.

Lv, Z., Tang, C., Zheng, Y., & Yang, X. (2025). Proactive Interaction of Artificial Intelligence Agents in Intelligent Assistive Systems: Mechanisms and Impacts. International Journal of Human–Computer Interaction, 1–21.

McCue, M., Khatib, R., Kabir, C., Blair, C., Fehnert, B., King, J., Spalding, A., Zaki, L., Chrones, L., Roy, A., & Kemp, D. E. (2023). User-Centered Design of a Digitally Enabled Care Pathway in a Large Health System: Qualitative Interview Study. JMIR Human Factors, 10(1), e42768.

Park, J. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–22.

Rathi, I., Taylor, S., Bergen, B. K., & Jones, C. R. (2024). GPT-4 is judged more human than humans in displaced and inverted Turing tests (No. arXiv:2407.08853). arXiv.

Schumann, A., & Lück, D. (n.d.). Better to ask online when it concerns intimate relationships? Survey mode differences in the assessment of relationship quality.

Seitl, M., Manuoglu, E., & Krám, T. (2024). Personal integrity and faking in the workplace: When competition matters. Current Psychology, 43(9), 7719–7730.

Takaffoli, M., Li, S., & Mäkelä, V. (2024). Generative AI in User Experience Design and Research: How Do UX Practitioners, Teams, and Companies Use GenAI in Industry? Proceedings of the 2024 ACM Designing Interactive Systems Conference, 1579–1593.

Teh, W. L., Abdin, E., P. V., A., Siva Kumar, F. D., Roystonn, K., Wang, P., Shafie, S., Chang, S., Jeyagurunathan, A., Vaingankar, J. A., Sum, C. F., Lee, E. S., van Dam, R. M., & Subramaniam, M. (2023). Measuring social desirability bias in a multi-ethnic cohort sample: Its relationship with self-reported physical activity, dietary habits, and factor structure. BMC Public Health, 23(1), 415.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. Psychological Bulletin, 133(5), 859–883.

Yang, R., Fu, M., Tantithamthavorn, C., Arora, C., Vandenhurk, L., & Chua, J. (2025). RAGVA: Engineering Retrieval Augmented Generation-based Virtual Assistants in Practice (No. arXiv:2502.14930). arXiv.