Redefining Research Crowdsourcing: Incorporating Human Feedback with LLM-Powered Digital Twins

Amanda Chan*
Princeton University
Princeton, New Jersey, USA
ac2921@princeton.edu

Catherine Di*
Princeton University
Princeton, New Jersey, USA
cathydi@princeton.edu

Joseph Rupertus*
Princeton University
Princeton, New Jersey, USA
joerup@princeton.edu

Gary Smith*
Princeton University
Princeton, New Jersey, USA
garysmith@princeton.edu

Varun Nagaraj Rao
Princeton University
Princeton, New Jersey, USA
varunrao@princeton.edu

Manoel Horta Ribeiro
Princeton University
Princeton, New Jersey, USA
manoel@princeton.edu

Andrés Monroy-Hernández
Princeton University
Princeton, New Jersey, USA
andresmh@princeton.edu

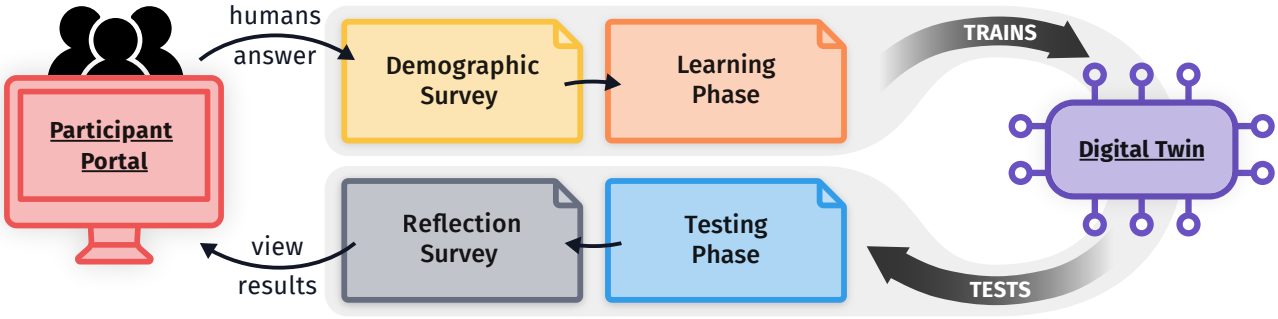

Figure 1: Participants complete surveys to train their digital twin and test its ability to answer questions in their place.

ABSTRACT

Crowd work platforms like Amazon Mechanical Turk and Prolific are vital for research, yet workers' growing use of generative AI tools poses challenges. Researchers face compromised data validity as AI responses replace authentic human behavior, while workers risk diminished roles as AI automates tasks. To address this, we propose a hybrid framework using digital twins, personalized AI models that emulate workers' behaviors and preferences while keeping humans in the loop. We evaluate our system with an experiment (n=88 crowd workers) and in-depth interviews with crowd workers (n=5) and social science researchers (n=4). Our results suggest that digital twins may enhance productivity and reduce decision fatigue while maintaining response quality. Both researchers and workers emphasized the importance of transparency, ethical data use, and worker agency. By automating repetitive tasks and preserving human engagement for nuanced ones, digital twins may help balance scalability with authenticity.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Information systems** → **Crowdsourcing**; **Collaborative and social computing systems and tools**.

KEYWORDS

digital twin, crowd work, AI uncertainty, MTurk, Prolific

*All four authors contributed equally to this research.

CHI EA '25, April 26-May 1, 2025, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/2025/04
https://doi.org/10.1145/3706599.3720269

ACM Reference Format:
Amanda Chan, Catherine Di, Joseph Rupertus, Gary Smith, Varun Nagaraj Rao, Manoel Horta Ribeiro, and Andrés Monroy-Hernández. 2025. Redefining Research Crowdsourcing: Incorporating Human Feedback with LLM-Powered Digital Twins . In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3706599.3720269

1 INTRODUCTION

Crowd work on online labor market platforms, such as Amazon Mechanical Turk (MTurk) and Prolific, is widely used in both industry and academia, supporting machine learning dataset creation as well as social science research examining human behavior [8, 27]. This paper focuses on crowd work in social science research, where the growing accessibility of Large Language Models (LLMs), such as ChatGPT and Gemini, has significantly reshaped the ecosystem [33]. Crowd workers increasingly utilize LLMs, perhaps due to the precarious nature of crowd work [8], often in ways that are difficult to detect [7, 28, 29].

These developments raise critical concerns for the researchers running the studies and the crowd workers completing them. For researchers, the growing use of LLMs by crowd workers threatens the validity of their research, which may reflect *AI* rather than *human* behavior. For crowd workers, LLM's growing capabilities threaten their livelihood by potentially reducing the demand for human labor in crowd work. Moreover, reliance on LLMs diminishes the value of their contributions as responses become less reliable and valuable for researchers.

Concerns about the quality of data generated by humans (often through crowd work) have led to a push for the creation of high-quality synthetic data [13]. Within this "synthetic framework," machine learning models and social science studies alike would be powered by synthetic outputs created by LLMs [2]. Perhaps unsurprisingly, this approach has been met with widespread criticism, as synthetic data generated by humans fails to capture the nuances of human text and behavior [20, 30].

To address these challenges, we propose a personalized, "hybrid," human- and AI-assisted framework for crowd working that benefits researchers and crowd workers. This approach introduces LLM-powered "digital twins" that accurately emulate workers' behaviors and preferences, responding only when confident, to preserve the integrity of individual contributions while keeping humans in the loop. We evaluated the system with 88 crowd workers, collecting feedback on their comfort and experience. Additionally, we interviewed five crowd workers and four social science researchers as key stakeholders to explore their experiences and perspectives on the system's potential impact.

Our study reveals key considerations for integrating AI into human-centered research while preserving research integrity and worker agency. Researchers were generally skeptical about AI's role in their studies, while crowd workers were less so. However, both agreed that the thoughtful implementation of digital twins could standardize response quality and reduce worker fatigue on repetitive tasks, making crowd work more accessible.

This work underscores the need for AI systems that augment, not replace, human judgment in research and data collection. It also highlights the importance of clear boundaries, transparency, and preserving human agency in AI-automated workflows. By prioritizing ethical and inclusive practices, our findings offer a framework for responsible AI integration that advances innovation while safeguarding human involvement.

2 RELATED WORK

2.1 Digital Twins and Personalized LLMs

Digital twins (DTs) are virtual models of natural, engineered, or social systems that replicate their structure, context, and behavior while continuously updating with data from their physical counterparts [1]. Although DTs have been applied in fields like civil engineering and medicine [3, 19], their use in crowd working to model human behavior and decision-making is still emerging. By prompting LLMs to emulate human preferences, personalized human DTs, or generative agents, have shown significant promise in accurately replicating human attitudes and behaviors [21, 22]. Past work also indicates that DTs can improve decision-making by aligning AI systems with human preferences [10]. In the context of crowd work, companies like Karya [15] and Qloo [24] are exploring AI-driven market research. Still, it is unclear whether they capture individual nuances or rely on broader demographic trends [12, 14]. These findings underscore the potential of digital twins when tailored to individual opinions, emphasizing the need for *personalized*, rather than generic, results. By focusing on individual-level LLM customization, this study seeks to improve the alignment and relevance of AI-assisted crowd work.

2.2 Worker-AI Collaboration in Crowd work

Integrating AI tools like LLMs into crowd work offers opportunities and challenges. While some researchers have demonstrated LLMs' ability to simulate human responses accurately [2, 33] and their potential for optimizing market research [6], others have cautioned that LLM-generated data may misrepresent social demographic groups (such as those defined by race, gender, or socioeconomic status), raising ethical concerns about authenticity and fairness [32]. Additionally, the widespread use of LLMs among crowd workers complicates the situation. Researchers have found that 33-46% of workers used LLMs for text summarization tasks, and even when explicitly asked to avoid LLM assistance, nearly half continued to use them covertly [28, 29]. This creates a dilemma: replacing humans with LLMs risks compromising diversity and authenticity, while traditional crowd work often involves generic, LLM-generated responses that fail to reflect individual human perspectives. Prior work has explored human-AI collaboration strategies to address these challenges that include: real-time AI feedback systems to improve crowd worker performance [32], dynamic task allocation to optimize task distribution between human and AI workers [17], new methods to evaluate AI-worker collaboration [14], and human verification of LLM outputs to improve annotation accuracy [31]. Building on these promising results, we propose and evaluate a digital twin system to enhance crowd worker productivity while preserving human agency and authenticity.

3 SYSTEM IMPLEMENTATION

We created a sample online research platform to simulate the use of digital twins to assist crowd workers in survey-answering tasks. The platform is a React application that interacts with OpenAI's GPT-4o model, an LLM, to create "digital twin" responses based on participant data.

Learning Phase Interface

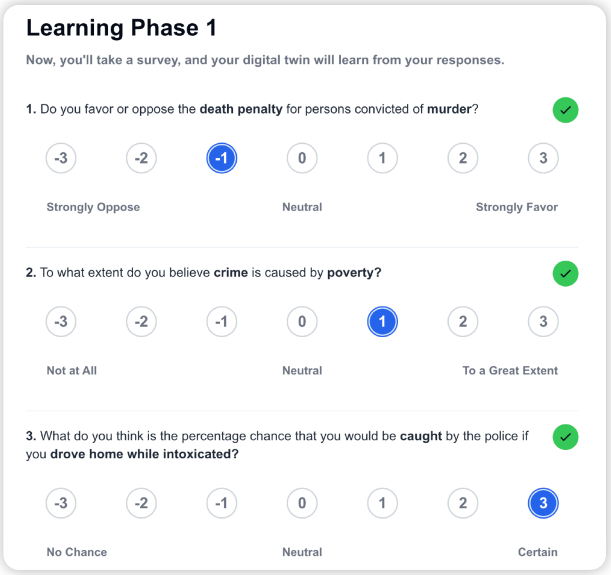

Figure 2: The learning survey interface design with Likert-scale questions.

Testing Phase Interface

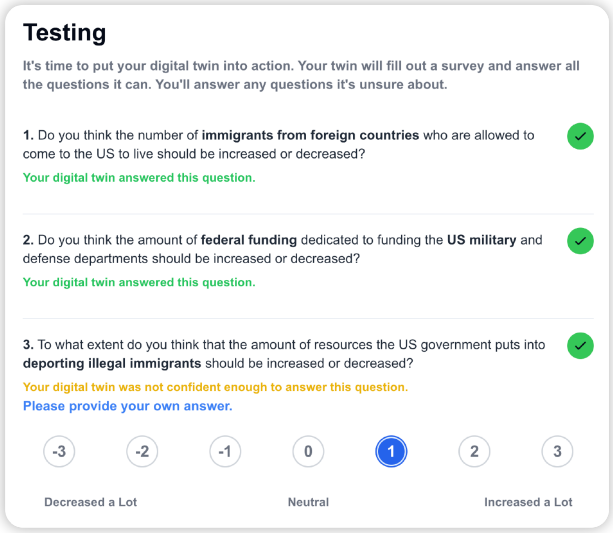

Figure 3: The testing survey interface with both digital twin and human responses.

3.1 Demographic Data Gathering

The system starts by asking crowd workers to answer 25 demographic questions (see Appendix C), including age, gender, education, location, and political views, to contextualize the digital twin's responses based on their own characteristics. These responses serve as foundational data for tailoring the digital twin's behavior to the individual participant.

3.2 Learning Phase

The crowd worker completes three learning surveys, each with 15 to 19 questions adapted from six social science surveys (see Appendix C), evaluated on a 7-point Likert scale from -3 to +3. These surveys, selected from Hewitt and colleagues' compilation of 70 U.S. studies [9], assess attitudes, beliefs, and perceptions on social, political, and ethical issues. After submitting the responses, the system prompts the LLM with the same questions to predict the participant's answers based on demographic data and prior responses (see Appendix D). The LLM's predictions are not shown to the user. Each phase iteratively refines the LLM's ability to simulate personalized answers.

3.3 Testing Phase

The system prompts the LLM with 43 additional survey questions from the same social science surveys as in the Learning Phase (see Appendix C) and instructs it to predict the participant's answers based on their demographic survey responses and learning survey responses. The OpenAI API response includes a list of probabilities corresponding to each token in the LLM's output. For each question, the system finds the token corresponding to the numerical Likert-scale prediction, extracts the logprob metric associated with it, and calculates the linear probability. This acts as a measure of the LLM's confidence in its answer. For a given question, if the confidence is above a certain threshold[1] (75%), the system accepts the answer and automatically fills out the survey question. The question will then appear to the participant as having been automatically completed by their "digital twin." If the confidence is below the threshold, the system rejects the LLM answer and defers to the participant to answer manually.

3.4 Reflection Survey

The participant independently answers all questions from the testing phase that were previously answered with high confidence (>75%) by the LLM.[2] The LLM's prediction is only visible after the participant gives their own answer. Finally, the participant answers 19 questions about the system (see Appendix C), providing feedback on the digital twin's usability, accuracy, trustworthiness in decision-making, and overall user experience and satisfaction.

4 METHODS

We recruited crowd workers (n = 88) to test our system's accuracy and usability and to understand their reactions to it. Further, we interviewed both crowd workers across the U.S. (n = 5) and social

[1]We empirically determined the 75% threshold for accepting digital twin answers as roughly the optimal point for maintaining accuracy while enabling the digital twin to answer enough questions to be useful.

[2]The reflection survey is intended to measure the LLM's accuracy in replicating individual participants' opinions, but it would not be included in a real system intended for use by crowd workers, since the intent is for the LLM to answer the questions automatically.

science researchers in the northeastern U.S. ($n = 4$) to better understand their perspectives on using AI to augment crowd work. This study was approved by Princeton University's Institutional Review Board (IRB# 17302), and all participants provided informed consent before participation.

4.1 Platform Study with Crowd Workers

We initially recruited 105 participants on Prolific, who were directed to our online platform. After excluding incomplete responses, the final sample size consisted of 88 participants. These individuals spent approximately 20–25 minutes completing surveys on the system and were compensated $5 for their time. Participants represented diverse backgrounds, all residing in the U.S. and originating from 32 different states. 54% reported that they rely on crowd work as their main source of income, while 46% have additional jobs. The majority of participants fell within the 26 to 45 age range. 58.1% identified as White, 22.1% as African American or Black, 11.6% as Hispanic/Latinx, and 5.8% as Asian. 54% identified as male and 46% identified as female.

The participants used the study platform (see Section 3), completing learning and testing surveys to train and evaluate their digital twin. The survey questions addressed six key topics: immigration/race, crime, health, politics, terrorism, and ethics. These topics were carefully selected to provide a comprehensive evaluation of the digital twin system's ability to replicate human responses across diverse and nuanced subject areas. All survey questions were sourced from original studies[3] that had been previously administered to human crowd workers. The survey questions were distributed such that 55% were part of the Learning Phase, while 45% were included in the Testing Phase. By covering all six topics in both phases, the digital twin system progressively familiarized itself with participant responses, mimicking the iterative learning process of humans.

4.2 Interviews with Crowd Workers and Researchers

We interviewed five crowd workers (recruited via Reddit and Prolific) and four social science researchers (recruited through personal connections). The crowd workers were all active users of Prolific and/or Amazon Mechanical Turk, and the researchers all used crowd working platforms in their research. Interviews were conducted virtually, lasting about one hour each. Crowd workers received $20 compensation, while researchers participated voluntarily.

The interviews explored participants' experiences with crowd working platforms and AI's influence on task quality, alongside a walkthrough of the digital twin system. Feedback focused on design, ethical concerns, privacy, and the long-term implications of AI in crowd work. Discussions compared AI-only, hybrid, and human systems in terms of cost, workload, authenticity, and fairness.

We audio-recorded and transcribed all interviews, then conducted thematic analysis using open coding. We coded the transcripts to identify recurring themes and patterns in user feedback. We iteratively refined these codes through discussion until reaching consensus, then grouped them into higher-level themes. This analysis revealed primary themes, including task-dependent trust and adoption, balancing efficiency with authenticity, privacy and data ownership concerns, quality control and oversight, and implementation requirements.

[3]Some phrasing was slightly modified to adapt to a Likert scale format.

5 RESULTS

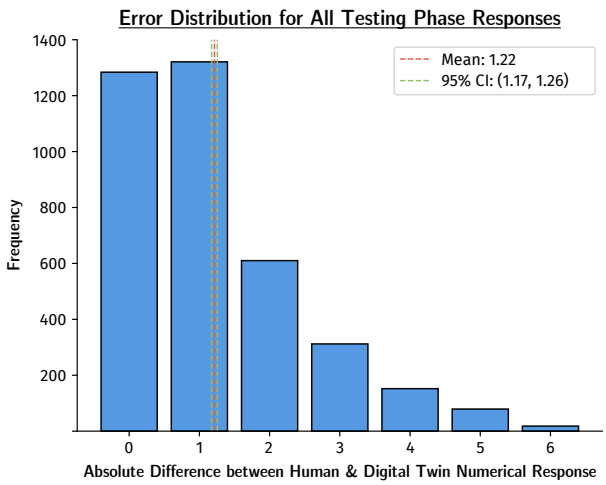

Figure 4: Distribution of absolute difference error for all test-phase responses (n = 3,784). An absolute difference of 0 means identical Human and Digital Twin Likert answers. Vertical lines show the mean error and 95% confidence interval.

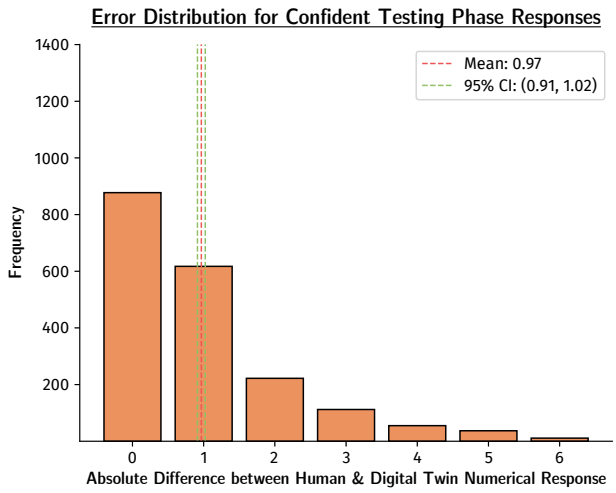

Figure 5: Distribution of absolute difference error for test-phase responses where the digital twin had high confidence (>75%) (n = 1,933). These questions were automatically answered, while others were deferred to humans. Vertical lines mark the mean error and 95% confidence interval.

5.1 Performance Metrics

5.1.1 Accuracy of Digital Twin Predictions. We measured the error of the digital twin's responses by computing the absolute difference between the digital twin's predictions and the participant's answers on a 7-point Likert scale. For questions the LLM answered independently, we compared its predictions with the participant's answers from the reflection survey. For questions the LLM deferred to the participant (it still made predictions in the background), we compared these predictions with the participant's actual answers. As shown in Figure 4, the mean error across all testing phase responses was 1.22. As shown in Figure 5, the mean error across all testing phase responses where the LLM answered in the user's place (i.e., it displayed high confidence) was 0.97. These results suggest that a digital twin is able to simulate an individual crowd worker's survey tasks with a high degree of accuracy.

5.1.2 Confidence of Digital Twin. We measured the confidence of the digital twin for each response based on the token probability of the LLM's numerical Likert scale prediction. The mean confidence across all question responses was 70.36%. Due to the 75% confidence threshold in our system, the system accepted the digital twin answer for 51.1% of questions on average across all testing phase responses and deferred the remaining 48.9% to human participants. For additional quantitative results, see Appendix A.

5.2 Crowd Worker Perspectives

5.2.1 Participants Recognize Accuracy but Emphasize the Need for Trust and Oversight. Participants acknowledged the digital twin's accuracy, with 60.7% of Prolific respondents reporting its recommendations often aligned with their choices and 31.5% noting they sometimes aligned (see Table 2 for more details). Participants described the system as typically accurate, deviating by only one point (P2, P4). However, trust varied: only 5.7% fully trusted it, while 25.0% expressed no trust. P2 remarked, *"If I worked with the same AI over and over, and it consistently chose things that were close to what I would choose, then I would probably trust it more."* Human oversight was seen as essential, with 42.0% of respondents preferring to monitor the system's decisions. P5 emphasized this, stating, *"I'd like to see what AI has written... it can make mistakes."*

5.2.2 Participants Value Efficiency but Highlight the Need for Human Involvement in Meaningful Tasks. The digital twin was praised for saving time, with 70.5% of respondents citing task automation as a key benefit and 52.3% valuing reduced decision fatigue. Participants appreciated AI for routine tasks and expressed frustration with repetitive, unstimulating work. P2 noted, *"It's hard when I'm dealing with the same kind of basic economic questions over and over again... I've probably seen them a thousand times."* Many favored a hybrid approach, allowing AI to handle routine tasks while humans focused on more engaging work. As P5 put it, *"AI could help with the repetitive questions... but human input is still needed, especially when the questions require thinking or writing."*

5.2.3 Participants Divided on Privacy, Prioritizing Authenticity Over Ethical Concerns. Privacy opinions were mixed: 36.7% of respondents were comfortable sharing data to improve accuracy, while 31.0% were uncomfortable. Similarly, opinions on crowd workers owning and monetizing digital twins were split, with 37.5% viewing it as harmful to privacy and 28.4% seeing it as beneficial. Many accepted data sharing as part of "life online," and prioritized authenticity over privacy (P2, P3, P4). As P3 explained, *"My only concern is the capabilities of the LLM to truly represent me."* Participants drew boundaries around sensitive data, such as personally identifiable information (PII), while appreciating AI's potential to enhance their work.

5.2.4 Improvements in Accuracy, Transparency, and Hybrid Models Key to Adoption. Key improvements included better prediction accuracy (67.0%), greater transparency in data usage (65.9%), and more control over privacy (56.8%). The flexibility to toggle between AI and human input was also prioritized. P4 highlighted the hybrid model's appeal, noting, *"If a study seems fun, let me figure this one out... workplace ones are dry, maybe a digital twin."* P5 added, *"If it's paying more, faster, and I can monitor it, I'd prefer the hybrid approach."*

5.2.5 Adoption and Recommendations Depend on Proven Benefits, Efficiency, and Human Oversight. Interest in digital twins was mixed: 48.9% expressed interest in future use, 28.4% were unsure, and 22.8% were unlikely to use the system. Participants emphasized the importance of proven benefits, such as improved accuracy (73.9%), cost efficiency (54.5%), and better privacy handling (50.0%). P1 expressed reservations, saying, *"I'm not comfortable with AI... but if it reduces scandalous activity, I'm fine with it."* Others were optimistic about a hybrid approach, with P1 affirming, *"hybrid is what's going to happen."* Human oversight remained critical for trust and adaptability.

5.3 Researcher Perspectives

5.3.1 Researchers Split on AI Detection and Trust in Diverse Crowd Work Tasks. Researchers expressed mixed opinions on detecting AI in crowd work. P7 highlighted challenges in identifying AI use in emotional response surveys, while P6 expressed trust in participants, stating, *"as a social scholar, I never thought about my respondents using AI."* For heterogeneous, non-text-based tasks, P8 noted, *"we cannot think of a way that you can easily adapt an AI to do the task."*

5.3.2 Researchers Prefer Human Responses but Acknowledge Cost Benefits of Human-AI Hybrid Approaches. Researchers strongly favored human responses for their authenticity and contextual richness. P9 stated, *"ideally, I prefer entirely human responses despite the challenges, as they offer insights into contextual and temporal factors."*, while P7 found hybrid systems *"a preferable option."* Others, like P8, viewed hybrid approaches as *"the worst of both worlds,"* believing that most tasks that could reliably use digital twins could just use AI entirely. Despite this, AI's cost and scalability benefits were acknowledged, with some willing to adopt AI if it was *"half the human cost"* (P6, P7). P9 observed, *"A hybrid approach, like your digital twin, combines efficiency with personalization but requires ongoing human feedback to remain effective."*

5.3.3 Transparency About AI Usage is Crucial for Research Integrity and Peer Review. Researchers stressed the importance of distinguishing AI from human responses for validity and peer review.

P6 advocated for *"clear information about whether a question is answered by [AI] or [human],"* allowing easier adjustments if AI usage is restricted. P9 emphasized transparency in the digital twin's training level for meaningful analysis, while P8 warned that excessive AI reliance could undermine research integrity. P6 also highlighted concerns about AI assumptions, noting *"these AI tools may be based on assumptions that people from the same school will hold similar opinions."*

5.3.4 Complex Tasks and Human Nuance Remain Beyond AI's Capabilities. Researchers highlighted AI's limitations in handling nuanced tasks, such as complex social issues and morally ambiguous topics (P9). P8 noted, *"It's difficult for AI to fully capture a person's complexity, as even a snapshot of someone is incomplete."* Variability in human opinions poses another challenge, with P8 emphasizing how responses shift based on phrasing and how digital twins may quickly become outdated as views evolve. P9 pointed out, *"you want to sometimes know the things that are influencing people's responses at that moment."* Additionally, AI often oversimplifies group representation, assuming, as P6 warned, *"people from similar backgrounds hold similar opinions,"* ignoring individual variation. These challenges highlight AI's struggle to reflect human perspectives' diverse, dynamic nature.

5.3.5 Clear Benchmarks and Standards Needed for Future AI Integration. Researchers foresee greater AI use in crowd work but remain skeptical about full automation. P9 noted the digital twin system might work for binary tasks but warned, *"for nuanced scales or exploratory research, small variations could matter,"* emphasizing the need to understand AI's limitations. P7 expressed concerns about unknowingly receiving predominantly AI-generated responses and stressed the importance of benchmarks, stating, *"95% accurate, for everyone, not 99% for some and 89% for others."* P8 agreed, adding, *"there must be a benchmark to which you compare the accuracy."* While researchers remain cautious to protect research validity, they acknowledge AI's potential when supported by clear standards and benchmarks.

6 DISCUSSION

Our study highlights key considerations for integrating AI in crowd work through digital twins. While our quantitative results show promising accuracy levels, researchers emphasized that numerical accuracy alone is insufficient, given the dynamic nature of human opinions and the influence of external events. This underscores the need for more sophisticated evaluation frameworks, such as routine sentiment checks or updates on personal opinions following significant social events.

We found a notable alignment between researcher and worker interests regarding data ownership and working conditions, with both groups emphasizing the importance of worker agency and control over digital representations. The effectiveness of digital twins appears highly task-dependent, with clear applications for repetitive elements like demographic questions while preserving human engagement for complex tasks requiring judgment. This natural division suggests opportunities for hybrid systems to improve working conditions and research integrity. Moving forward, aligning AI capabilities with worker needs and ethical considerations will be critical to ensuring that digital twins support sustainable and fair crowd work ecosystems.

Privacy and data usage are central to the ethical implementation of this study. To ensure transparency and participant awareness, all demographic information in this study was collected with informed consent and securely stored in compliance with human subjects research guidelines. While leveraging demographic data to tailor digital twins can enhance personalization, it also introduces risks related to data security and misuse; thus, strict safeguards were implemented to prevent unauthorized access and ensure data confidentiality. Inclusive demographic questions (see Appendix C) were also used to promote fairness and recognition of individual diversity. Moving forward, research should prioritize methods that avoid reinforcing assumptions associated with demographic categories.

The ethical implications of integrating AI into crowd work also extend beyond data privacy to include concerns about the potential replacement of human judgment, along with its impact on worker livelihood and job security. While digital twins can alleviate repetitive tasks and reduce cognitive load, their widespread use may reduce demand for human input, affecting income stability for crowd workers. This emphasizes the need to deploy AI as a collaborative tool that enhances human capabilities rather than replacing them, preserving opportunities for meaningful and creative tasks that require human judgment.

7 LIMITATIONS AND FUTURE WORK

From a technical perspective, our implementation using general-purpose LLMs (GPT-4o) presents both opportunities and limitations. The inherent biases in historical training data and the potential inability to capture responses to current events suggest the need for alternative approaches. However, resource constraints present practical challenges when training specialized models. Future work might explore creative solutions like enhanced prompting strategies or hybrid architectures that balance accuracy with resource efficiency.

Additionally, several limitations of our study warrant consideration, including sample size constraints and potential self-selection bias in our interview participants. Differences in research methodologies among participating researchers could also influence attitudes toward AI integration. For instance, one psychology researcher expressed little concern about AI use (P8), given that their research relied on non-text-based survey designs. Future research should investigate digital twins' applicability to tasks involving free-text responses, subjective assessments, and contextual interpretation, exploring how AI can effectively support diverse research needs.

Finally, the study's focus on Likert-scale tasks limits the generalizability of our findings to more complex or nuanced annotation work. Future research should focus on expanding digital twin capabilities beyond Likert scales, developing flexible systems for AI control, implementing adaptive oversight mechanisms, and making crowd work more accessible to diverse populations. While digital twins demonstrate promise to enhance the efficiency and quality of crowd work, their successful implementation depends on balancing technical capabilities with the evolving needs of researchers and preserving the integrity of human-centered research.

8 CONCLUSION

Our study highlights the potential of hybrid AI-human systems, particularly digital twins, to revolutionize crowd work by balancing efficiency and authenticity. While the proposed framework automates repetitive tasks and maintains research integrity, it underscores the importance of human agency and task-specific AI deployment. Both crowd workers and researchers acknowledge the benefits of such systems but emphasize the need for transparency, trust, and oversight. Our findings demonstrate that aligning worker and researcher interests through clear benchmarks, flexible AI usage, and robust privacy safeguards can lead to ethical and inclusive AI integration. Future work should expand on these foundations, addressing limitations in model adaptability, task diversity, and system scalability to ensure broader applicability and fairness in AI-augmented research.

REFERENCES

[1] AIAA. 2020. Digital Twin: Definition & Value – An AIAA and AIA Position Paper. https://www.aia-aerospace.org/publications/digital-twin-definition-value-an-aiaa-and-aia-position-paper/

[2] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. Out of One, Many: Using Language Models to Simulate Human Samples. https://doi.org/10.1017/pan.2023.2

[3] Nandeesh Babanagar, Brian Sheil, Jelena Ninić, Qianbing Zhang, and Stuart Hardy. 2025. Digital twins for urban underground space. *Tunnelling and Underground Space Technology* 155 (Jan. 2025), 106140. https://doi.org/10.1016/j.tust.2024.106140

[4] Daniel Silverman, Daniel Kent, and Christopher Gelpi. 2020. Can Factual Misperceptions be Corrected? An Experiment on American Public Fears of Terrorism. (June 2020). https://osf.io/a7uk3/ Publisher: OSF.

[5] Michael Davern, Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. 2024. General Social Survey 1972-2024. gssdataexplorer.norc.org

[6] Diana-Elena Drăghici, Andreea Orîndaru, Mihaela Constantinescu, and Alina Zelezneac. 2023. Revolutionizing Marketing Research Through AI: comprehensive review of the past, present, and future. *Journal of Emerging Trends in Marketing and Management* I, 1 (May 2023), 39–45. http://www.etimm.ase.ro/RePEc/aes/jetimm/2023/ETIMM_V01_2023_73.pdf Publisher: The Bucharest University of Economic Studies Publishing House.

[7] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.

[8] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

[9] Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting Results of Social Science Experiments Using Large Language Models. (Aug. 2024).

[10] Yijun Huang, Jihan Zhang, Xi Chen, Alan H. F. Lam, and Ben M. Chen. 2024. From Simulation to Prediction: Enhancing Digital Twins with Advanced Generative AI Technologies. In *2024 IEEE 18th International Conference on Control & Automation (ICCA)*. 490–495. https://doi.org/10.1109/ICCA62789.2024.10591881 ISSN: 1948-3457.

[11] Jennifer L. Hughes, Abigail A. Camden, Tenzin Yangchen, Gabrielle P. A. Smith, Melanie M. Domenech Rodríguez, Steven V. Rouse, C. Peeper McDonald, and Stella Lopez. 2022. Guidance for Researchers When Using Inclusive Demographic Questions for Surveys: Improved and Updated Questions. *Psi Chi Journal of Psychological Research* 27, 4 (2022), 232–255. https://doi.org/10.24839/2325-7342.JN27.4.232

[12] IBM. 2021. What Is a Digital Twin? https://www.ibm.com/topics/what-is-a-digital-twin

[13] Business Insider. 2024. The AI world's most valuable resource is running out, and it's scrambling to find an alternative: 'fake' data. (2024). https://www.businessinsider.com/ai-synthetic-data-industry-debate-over-fake-2024-8 Accessed: 2025-01-19.

[14] Tomoya Kanda, Hiroyoshi Ito, and Atsuyuki Morishima. 2022. Efficient Evaluation of AI Workers for the Human+AI Crowd Task Assignment. In *2022 IEEE International Conference on Big Data (Big Data)*. 3995–4001. https://doi.org/10.1109/BigData55660.2022.10020844

[15] Karya. [n. d.]. Karya Institute. https://institute.karya.in/

[16] Katherine McCabe. 2019. Public Opinion and Attributions for Health Care Costs. (May 2019). https://osf.io/3pcdm/ Publisher: OSF.

[17] Masaki Kobayashi, Kei Wakabayashi, and Atsuyuki Morishima. 2021. Human+AI Crowd Task Assignment Considering Result Quality Requirements. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9 (Oct. 2021), 97–107. https://doi.org/10.1609/hcomp.v9i1.18943

[18] Maureen Craig. 2017. Racial Majority & Minority Group Members' Psychological and Political Reactions to Minority Population Growth. (Nov. 2017). https://osf.io/sazxn/ Publisher: OSF.

[19] Divya Nagaraj, Priya Khandelwal, Sandra Steyaert, and Olivier Gevaert. 2023. Augmenting digital twins with federated learning in medicine. *The Lancet. Digital Health* 5, 5 (May 2023), e251–e253. https://doi.org/10.1016/S2589-7500(23)00044-4

[20] Vishakh Padmakumar and He He. 2024. Does Writing with Language Models Reduce Content Diversity? arXiv:2309.05196 [cs.CL] https://arxiv.org/abs/2309.05196

[21] 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 (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3586183.3606763

[22] 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. https://doi.org/10.48550/arXiv.2411.10109 arXiv:2411.10109 [cs].

[23] Mark Peffley and Jon Hurwitz. 2007. Persuasion and Resistance: Race and the Death Penalty in America. *American Journal of Political Science* 51, 4 (2007), 996–1012. https://doi.org/10.1111/j.1540-5907.2007.00293.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5907.2007.00293.x.

[24] Qloo. [n. d.]. Qloo | Develop Personalized Experiences With Taste AI. https://www.qloo.com

[25] Rebecca Bucci. 2023. Accounting for the Correlation between Perceived Risks and Rewards to Crime. https://osf.io/4fsg5/

[26] Rochelle Terman. 2020. Human Rights Shaming, Compliance, and Nationalist Backlash. (Jan. 2020). https://osf.io/q8ra3/ Publisher: OSF.

[27] Matthew J Salganik. 2019. *Bit by bit: Social research in the digital age*. Princeton University Press.

[28] Veniamin Veselovsky, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2023. Prevalence and prevention of large language model use in crowd work. https://doi.org/10.48550/arXiv.2310.15683 arXiv:2310.15683 [cs].

[29] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. https://arxiv.org/abs/2306.07899v1

[30] Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2024. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908* (2024).

[31] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, 2024-05-11). ACM, 1–21. https://doi.org/10.1145/3613904.3641960

[32] Yunlong Wang, Priyadarshini Venkatesh, and Brian Y Lim. 2022. Interpretable Directed Diversity: Leveraging Model Explanations for Iterative Crowd Ideation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–28. https://doi.org/10.1145/3491102.3517551

[33] Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Bill Guo, Sireesh Gururaja, Tzu-Sheng Kuo, et al. 2023. Llms as workers in human-computational algorithms? replicating crowdsourcing pipelines with llms. *arXiv preprint arXiv:2307.10168* (2023).

A ADDITIONAL RESULTS

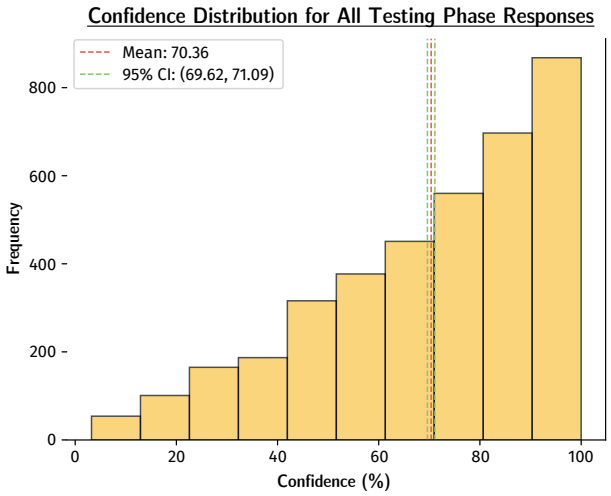

Figure 6: The distribution of digital twin confidence for all question responses from the testing phase.

B INTERVIEW PROTOCOL

(1) **Introduction and Consent (5 minutes):**
- Researchers introduce themselves and explain the purpose of the study.
- Participants confirm they have read and signed the consent form.

(2) **Prototype Use (20 minutes):**
- Participants use our prototype system by answering a series of surveys and interacting with a "digital twin" that assists them in answering questions.
- Participants evaluate the digital twin based on perceived accuracy and alignment with their real preferences. They also answer feedback questions about the system.

(3) **Discussion (35 minutes):**
- Questions focus on the participants' thoughts on integrating digital twins into their workflows, their perceptions of the hybrid approach's validity, and potential barriers to adoption.

(4) **Conclusion (5 minutes):**
- Summarize key discussion points and ask if participants have additional thoughts.
- Thank participants for their time.

C SURVEY QUESTIONS

The following list provides a breakdown of the sources referenced for each survey question. The "Supplementary Materials" document includes the complete list of survey questions included in the study. Many of these questions are either directly taken from or inspired by established research studies, as detailed below. The sources were identified through the work of Hewitt and colleagues, who compiled 70 nationally representative U.S. studies [9].

Below each *free-response* question, participants saw the following:

> "Please do not include any directly or indirectly personally identifiable information, such as your name or address."

Demographics

Source: Guidance for Researchers When Using Inclusive Demographic Questions for Surveys: Improved and Updated Questions [11]
- Demographic Survey: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 21, 22, 24

Source: 2022 GSS Cross-Section Questionnaire in English [5]
- Demographic Survey: 11, 12, 13, 14, 15, 16, 17, 18 19, 20, 23, 25

Race and Death Penalty

Source: Persuasion and Resistance: Race and the Death Penalty in America [23]
- Learning Survey 1: Questions 1, 2
- Learning Survey 2: Questions 3, 4
- Learning Survey 3: Questions 3, 4
- Testing Phase: Questions 6, 7

Crime Risks vs. Reward

Source: Accounting for the Correlation between the Perceived Risks and Rewards to Crime [25]
- Learning Survey 1: Questions 3, 4, 5, 6
- Learning Survey 2: Questions 5, 6, 7, 8
- Learning Survey 3: Questions 5, 6, 7, 8
- Testing Phase: Questions 8, 9, 10, 11, 12, 13, 14, 15

Healthcare Costs

Source: Public Opinion and Attributions for Health Care Costs [16]
- Learning Survey 1: Questions 7, 8
- Learning Survey 2: Questions 9, 10, 11
- Learning Survey 3: Questions 9
- Testing Phase: Questions 16, 17, 18, 19, 20, 21

Human Rights Backlash

Source: Human Rights Shaming, Compliance, and Nationalist Backlash [26]
- Learning Survey 1: Questions 9, 10, 11, 12
- Learning Survey 2: Questions 12, 13, 14
- Learning Survey 3: Questions 10, 11
- Testing Phase: Questions 22, 23, 24, 25, 26, 27, 28, 29, 30, 31

Terrorism Fear Correction

Source: Can Factual Misperceptions be Corrected? An Experiment on American Public Fears of Terrorism [4]
- Learning Survey 1: Questions 13, 14, 15, 16
- Learning Survey 2: Questions 15, 16, 17, 18
- Learning Survey 3: Questions 12, 13, 14, 15
- Testing Phase: Questions 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43

Testing Results	Mean Absolute Difference (95% CI)	Mean Confidence (%) (95% CI)	Total Number of Responses
All Testing Responses	1.22 (1.17–1.26)	70.36 (69.62–71.09)	3,776
All Testing Responses Independently Answered by Digital Twin	0.97 (0.91–1.02)	88.82 (88.47–89.16)	1,931 (51.1%)
All Testing Responses Deferred and Answered by Human	1.47 (1.42–1.53)	51.03 (50.24–51.82)	1,845 (48.9%)

Table 1: Testing results with mean absolute difference, mean confidence, and total number of responses.

Racial Minority Growth

Source: Racial Majority & Minority Group Members' Psychological and Political Reactions to Minority Population Growth [18]

- Learning Survey 1: Questions 17, 18, 19
- Learning Survey 2: Questions 1, 2
- Learning Survey 3: Questions 1, 2
- Testing Phase: Questions 1, 2, 3, 4, 5

D LLM PROMPT

This prompt was submitted to GPT-4o via OpenAI's API after each learning survey and testing survey submission:

```
For context, we are currently running a study on how well LLM-powered digital twins are able to answer crowdsourcing surveys based on human input. Our aim is for the digital twin to effectively and accurately mimic the human crowd worker.
You are the digital twin model in this study. I will provide you with the human crowd worker's demographic information and preferences, and your task is to respond to questions as closely as you can to the human crowd worker's preferences in the strict format outlined below. Do not use any outside knowledge other than what I've given you on how the human crowd worker thinks. This task is very important. Do this to the best of your ability. If you do not know the answer to a question based on the human crowd worker's input, output "$".
All references to the word "human" mean the specific "human crowd worker" that you are trying to mimic. There is only one human you need to mimic.
For the questions in this survey, you will be given the following information as context: 1. The demographic info of the human that you are trying to mimic. 2. The human's answers to the prior survey before this survey. 3. The LLM-powered digital twin's answers to the same survey that the human answered. The purpose of this study is for the LLM-powered digital twin to learn from the human's responses to better mimic the human crowd worker. Therefore, the human will answer the same survey that you answer, and you need to try to learn from the human's responses to better understand their opinions and preferences. Whenever the answers from (2) and (3) do not match up, always prioritize the human's answer.
Your task is split into two steps:
Step 1: Write a short paragraph summarizing the key points of the human's demographic information and preferences. IMPORTANT: If you did not receive any demographic info, for step 1, simply output "Did not receive demographic information."
Step 2: Using all of the context information given to you as well as the paragraph you wrote in Step 1, answer the survey questions given to you to the best of your ability following the response format below.
To clarify, the only things you will output back are the paragraph from Step 1 and the numerical answers from Step 2.
How to answer survey questions: Do your best to predict what the human would choose for each question on a scale of 1 to 7: - 1 = Strongly Disagree (or appropriate extreme) - 7 = Strongly Agree (or appropriate extreme) - Intermediate values (2-6) represent varying degrees between these extremes.
Please format your responses in JSON format as follows:
Sample response structure: { "step1": "Your paragraph here", "step2": { "1": "1", "2": "5", "3": "3", "4": "2", "5": "4", "6": "5", "7": "1", "8": "6", "9": "2", "10": "1" } }
```

Following these instructions, a JSON object containing the user's demographic data and all numerical answers from previous learning surveys by both the human and the LLM is submitted.

Question	Option 1	Option 2	Option 3	Option 4	Option 5
How well do the recommendations from the digital twin align with your choices?	1.1% Always align	60.7% Often align	31.5% Sometimes align	6.7% Rarely align	0.0% Never align
Do you think systems like this could improve or harm the conditions for crowd workers?	9.1% Strongly improve	34.1% Somewhat improve	19.3% No significant impact	27.3% Somewhat harm	10.2% Strongly harm
How do you feel about the concept of using digital twins for survey-answering tasks?	23.0% Very excited	28.7% Somewhat excited	18.4% Neutral	16.1% Somewhat skeptical	13.8% Very skeptical
How comfortable are you with sharing personal data to improve the accuracy of your digital twin?	19.5% Very comfortable	17.2% Somewhat comfortable	19.5% Neutral	17.2% Somewhat uncomfortable	13.8% Very uncomfortable
How would allowing crowd workers to own and monetize their digital twins affect your data privacy concerns?	9.1% Strongly improve	19.3% Somewhat improve	34.1% No significant impact	29.5% Somewhat harm	8.0% Strongly harm
Would you accept slightly lower pay if the platform reduced your workload by using AI to handle repetitive tasks?	10.2% Yes, definitely	21.6% Yes, to some extent	19.3% Not sure	20.5% No, unlikely	20.5% No, definitely not
To what extent would you trust a digital twin to make decisions on your behalf?	5.7% Fully trust it to make decisions for me	42.0% Trust it with some decisions, but prefer oversight	27.3% Trust it only with minor decisions	25.0% Don't trust it to make any decisions	N/A
Out of everything in this system, what stood out to you the most?	33.0% Ease of use	31.8% Alignment with my choices	11.4% Ethical and privacy concerns	15.9% Accuracy of predictions	5.7% Potential benefits for crowd workers
In what areas has (or do you expect) the digital twin to help you the most? (select all that apply)	70.5% Saving time by automating tasks	52.3% Reducing decision fatigue by providing recommendations	52.3% Gaining insights about personal habits or behavior	23.9% Improving my physical or mental well-being	N/A
What improvements would make you more likely to use a digital twin regularly? (select all that apply)	67.0% Better accuracy of predictions	65.9% More transparency about how data is used	56.8% Greater control over data collection and privacy	37.5% Improved user interface and ease of use	42.0% Stronger alignment with my personal values
Would you use a digital twin like this in the future?	14.8% Yes, definitely	34.1% Yes, likely	28.4% Maybe, unsure	14.8% No, unlikely	8.0% No, definitely not
Would you recommend this system to others in your field?	18.4% Yes, highly recommend	20.7% Yes, recommend with some reservations	32.2% Neutral	19.5% No, would not recommend	9.2% No, strongly against recommending
What would increase your confidence in adopting or recommending this platform? (select all that apply)	42.0% Clearer explanation of how it works	54.5% Demonstrated cost and time efficiency	73.8% Evidence of improved accuracy over alternatives	50.0% Better handling of ethical or privacy concerns	N/A

Table 2: Summary of Reflection Survey Responses