Journal of Artificial Intelligence Practice (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2026.090119
ISSN 2371-8412 Vol. 9 Num. 1

Silicon Samples: A Review and Outlook on Large Language Models Simulating Human Respondents

Yuyi Zhang

School of Business Administration, Guizhou University of Finance and Economics, Guiyang, Guizhou, 550025, China

Keywords: Large language models; Silicon samples; Algorithmic Fidelity; Psychometrics and Machine Psychology; Consumer and Market Research; Digital Twins; Political Opinion Simulation

Abstract: The use of large language models (LLMs) to simulate human respondents (silicon samples), as an emerging research topic, is attracting growing scholarly attention. By reviewing 25 representative studies from both domestic and international literature, this paper offers a definition of LLM-based simulation of human samples and examines its foundations, characteristics, and purposes. It summarizes the applications of this method across four major domains—psychometrics and machine psychology, consumer and market research, experimental simulation and digital twin construction, and the simulation of political opinion and social sentiment. It further reviews the existing evidence regarding the method's validity, along with the attendant debates, along three dimensions: psychometric validity, group fidelity and context dependence, and ethical and epistemic justice risks. Finally, the paper synthesizes a research framework for LLM-based simulation of human samples and discusses future research directions, with the aim of providing a reference for research and applied practice in this field.

1. Introduction

The rapid development of large language models (LLMs) is changing how the social sciences and consumer research obtain human data. Traditional surveys and experiments remain central to the study of attitudes, preferences, and behavior, but representative recruitment is costly and time-consuming, response rates are declining, and some populations remain difficult to reach [1]. Trained on large-scale human text, LLMs encode linguistic, psychological, and sociocultural patterns and can be conditioned with demographic or role information to respond as particular individuals or groups [2]. This has encouraged the use of LLM-generated “silicon samples” to supplement or partially replace human participants, including emerging applications in product research and digital-twin datasets [3]. Evidence, however, is mixed: some studies report substantial algorithmic fidelity [2], whereas others identify distortion of identity groups [4], cross-cultural deviations [5], and tensions between safety alignment and fidelity [6]. Clarifying the concept, applications, and validity boundaries of this method is therefore necessary.

This study first uses three sets of keywords—"large language model / generative AI (LLM/generative AI)," "human sample / respondent (human sample/respondent)," and "silicon

sampling / algorithmic fidelity / digital twin (silicon sampling/algorithmic fidelity/digital twin)"—in cross-combination as subject terms to search foreign-language core databases such as Web of Science. Results were screened by reading titles, keywords, and abstracts, and purely technical algorithmic improvements and other loosely related literature were excluded, yielding an initial set of 16 articles. A "snowball" search and screening based on their references was then conducted, ultimately producing 25 representative articles that met the criteria, drawn from top international journals in marketing, psychology, political science, and sociology, including the *Journal of Marketing*, *Marketing Science*, *Political Analysis*, *Nature Machine Intelligence*, and *PNAS*. This review reveals that current research tends to focus on a single disciplinary task (such as personality measurement, brand perception, or political sentiment) in a fragmented manner, lacking a systematic cross-disciplinary synthesis, which hinders a clear grasp of the applicable boundaries and risks of this emerging method.

Accordingly, this paper addresses four questions: how LLM-based simulation of human samples should be defined, where it is applied, how its validity and limitations are evaluated, and how the field may develop. The review identifies conditional generation as its technical foundation; anthropomorphism, scalability, and controllability as key characteristics; and the replacement or supplementation of human participants as its principal purpose. Applications concentrate in psychometrics and machine psychology, consumer and market research, experimental simulation and digital twins, and political opinion and social sentiment. Validity remains conditional and contested across psychometric performance, group fidelity and context dependence, and ethical and epistemic justice.

2. The Connotation of LLM-Based Simulation of Human Samples

Existing literature has not yet established a unified definition of LLM-based simulation of human samples. Related concepts emphasize algorithmic fidelity, “silicon samples,” or “digital doubles,” respectively highlighting the model’s capacity to reproduce subgroup patterns, generate synthetic datasets, or stand in for respondents [2, 7, 8]. Synthesizing these perspectives, this paper defines the method as a research paradigm in which LLMs are conditioned with demographic, role, or individual information to generate synthetic responses for surveys, psychometric measurement, or experiments, thereby replacing or supplementing human participants. Its connotation can be understood through its foundation, characteristics, and purpose.

2.1. Foundation: The Conditional Generation Logic of LLMs

The method rests on LLMs’ capacity to learn psychological and sociocultural regularities from large-scale human-generated text. Such corpora contain linguistic habits, values, and group-specific psychological patterns, consistent with the lexical hypothesis that socially important individual differences become encoded in everyday language [9]. Human-like responses generated by LLMs therefore derive from statistical learning over these linguistic patterns.

At the operational level, simulation relies on conditional generation. Researchers embed demographic characteristics, personality descriptions, or interview information in prompts so that the model responds from a designated identity perspective [2, 10]. The process can be summarized as input, generation, and calibration: profile information is structured as a prompt [3]; the model produces natural-language or constrained responses [11]; and output stability and variability are managed through parameters such as temperature and repeated generation [12].

2.2. Characteristics: Anthropomorphism, Scalability, and Controllability

LLM-based simulation has three main characteristics. First, anthropomorphism: models can reproduce not only human-like language but also measurable personality patterns and, in some settings, deeper psychological effects such as cognitive dissonance [13]. Second, scalability: automated simulation can generate large numbers of responses rapidly and at much lower cost than conventional recruitment [8]. Third, controllability: prompt design, parameter settings, fine-tuning, and alignment can systematically shape response stability, personality expression, and downstream behavior [9, 11].

2.3. Purpose: Replacing or Supplementing Human Participants

Its core purpose is to replace or supplement human participants in order to improve efficiency and accessibility. LLMs can reduce the cost and time required for pretesting and early psychometric evaluation [7, 12] and can help researchers explore populations that are difficult to recruit directly [8]. Their most appropriate role is therefore increasingly framed as collaboration with, rather than wholesale substitution for, human participants and researchers, with human judgment retained for theory development, hypothesis testing, and interpretation [14].

3. Application Domains of LLM-Based Simulation of Human Samples

Current applications span several stages of empirical research and reflect the method's advantages in accessibility, scale, and reach [8, 14]. Synthesizing the literature, this paper groups the main uses of LLM-based human-sample simulation into four domains: psychometrics and machine psychology, consumer and market research, experimental simulation and digital twin construction, and political opinion and social sentiment.

3.1. Psychometrics and Machine Psychology

Psychometric scale development and validation traditionally depend on cross-cultural human samples and can be costly and slow [15]. LLMs provide a rapid, low-cost supplementary environment for preliminary item generation, testing, and model-based psychological analysis.

LLMs can support automatic questionnaire-item generation and preliminary validity screening. Virtual respondents with trait-response mediators have been used to test construct validity, addressing the tendency of earlier automated approaches to emphasize reliability more than whether items measure the intended construct [15]. Personality measurement has also become a major line of inquiry. GPT-4o was evaluated on the Eysenck Personality Questionnaire across six languages, while GPT-3.5, GPT-4, and GPT-4o were compared under different temperature settings; the results showed partial reproduction of real-sample patterns, including demographic and disciplinary differences [12]. A broader framework covering eighteen models from five model families further found that model personality traits can display measurable construct properties, be deliberately shaped along target dimensions, and influence downstream behavior [11]. Research has also moved beyond trait measurement: GPT-4o exhibited a cognitive-dissonance effect moderated by free choice in preregistered trials [13], and LLM-generated questionnaire responses have been examined for whether they reproduce the internal structure of self-regulated learning [16]. Overall, these studies suggest that LLMs are evolving from simple questionnaire-answering tools into supplementary psychometric platforms for item validation, personality analysis, and the study of machine psychological processes.

3.2. Consumer and Market Research

Traditional market research depends heavily on consumer surveys and interviews, which can be expensive and time-consuming, particularly for smaller firms [17]. LLMs offer a lower-cost supplementary tool for exploratory market analysis.

LLMs have been used for specific consumer-research tasks including brand perception and segmentation. Model-generated brand-similarity and attribute-rating data have shown potential for perceptual analysis without full-scale human surveys [18]. LLM-based text embeddings have also been used to connect open-ended and multiple-choice questionnaire responses and to construct consumer-profile clusters, with an empirical application to market segmentation [19]. In addition, intertemporal-choice research has tested whether LLMs can reproduce human patterns of temporal discounting [20]. Beyond individual tasks, scholars have assessed LLMs as collaborators across multiple stages of the marketing research value chain [14]. Comparisons between four mainstream LLMs and a survey of 1,003 consumers show that simulation reliability tends to decline as questions become more specific to products and contexts [17], consistent with broader reviews that emphasize both the opportunities and limitations of silicon samples in consumer research [7]. Related studies extend the method to sensory evaluation [21] and to service-failure and recovery research comparing human and synthetic data [22]. Taken together, current evidence indicates stronger value for exploratory analysis, problem definition, and early-stage insight generation than for replacing human respondents in context-specific consumer judgments.

3.3. Experimental Simulation and Digital Twin Construction

Experimental replication is often constrained by recruitment, time, and geography. LLMs create new opportunities for low-cost replication, experimental pretesting, and the construction of digital representations of respondents.

LLMs have been used to replicate published experiments and to support new experimental designs. GPT-based simulations have reproduced classic tourism experiments, with robustness checks used to assess result stability [23]. ChatGPT-4 and LLaVA have also been evaluated for predicting human preferences for natural landscapes, extending simulation to visual and environmental assessment [24]. At the design stage, the LOLA algorithm combines LLM predictions with online learning to reduce cold-start problems and identify promising content variants under limited traffic [25]. Digital-twin construction represents a related application. The Twin-2K-500 dataset was built from more than 2,000 respondents answering over 500 questions; its evaluation shows that individual-level prediction remains limited, whereas aggregate population distributions can be reproduced more effectively [3]. Interview-informed role prompting has likewise been used to compare AI-generated and human questionnaire responses, illustrating a digital-twin approach that combines richer respondent information with scalable simulation [10]. These applications show that LLMs can support replication, experimental optimization, and population-level digital representation, although individual-level fidelity remains a key constraint.

3.4. Simulation of Political Opinion and Social Sentiment

Rising survey costs and declining response rates have also motivated the use of LLMs in research on public opinion and social attitudes [1]. This domain provides an important test of whether algorithmic fidelity can extend from general language behavior to socially and culturally situated attitudes.

Foundational work suggested that appropriately conditioned language models can approximate the attitudinal and behavioral patterns of specific subpopulations [2]. Subsequent studies have qualified

this conclusion. Synthetic survey data may not reliably recover human means, variances, and correlation structures [1]; fidelity can fall substantially for culturally embedded or sensitive issues in non-Western contexts [5]; and alignment can increase refusals and normative commentary while distorting attitude distributions [6]. Ethical analyses further warn that synthetic stand-ins may flatten or misrepresent heterogeneous identity groups [4]. Thus, LLMs may reproduce some aggregate public-opinion patterns, but their use as substitutes for human respondents requires strong contextual and subgroup-level validation.

4. Evidence and Limitations of the Validity of LLM-Based Simulation of Human Samples

The value of LLM-based simulation ultimately depends on the validity of the data it generates. The input, generation, and calibration stages correspond to three major sources of concern: psychometric reliability and validity, group fidelity and context dependence, and ethical or epistemic risks associated with alignment and representation. The following review therefore evaluates validity along these three dimensions.

4.1. Psychometric Validity

Psychometric evidence shows that reliability and validity vary substantially across constructs and settings. For example, GPT-4o has shown high internal consistency on some personality dimensions but poor reliability on others, indicating that strong reliability cannot be assumed across scales [12]. More importantly, consistency does not by itself establish construct validity; automated questionnaire research often tests response stability without demonstrating that an item measures the intended construct [15]. Results are also sensitive to parameter standardization, prompt engineering, demographic conditioning, and acquiescence bias [9, 16]. Psychometric validity should therefore be treated as task- and configuration-dependent rather than as a general property of LLM-generated data.

4.2. Group Fidelity and Context Dependence

Evidence on group fidelity likewise combines promise with important qualifications. Early work found that conditioned models could reproduce demographic variation in human attitudes and behaviors [2], but replication studies identified weaknesses in recovering sentiment distributions and conditional correlations [1]. Cross-cultural evidence further shows higher agreement on low-sensitivity issues than on culturally embedded or contested topics [5]. Alignment introduces an additional trade-off: although it can improve compliance and controllability, it may raise refusal rates and shift attitude distributions away from those observed in human samples [6]. Algorithmic fidelity is therefore conditional on the task, issue sensitivity, model version, cultural context, and alignment strategy.

4.3. Ethical and Epistemic Justice Risks

Beyond statistical validity, LLM-based substitution raises ethical and epistemic concerns. Models may compress the internal diversity of identity groups into stereotyped “outsider” representations, thereby misportraying lived experience [4]. Because simulation is inexpensive and easily scalable, it may also encourage researchers to speak for groups without engaging actual participants, increasing the risk of misuse [8]. Efficiency gains must therefore be weighed against the possibility that synthetic data reproduce or amplify existing social biases and exclude the perspectives they are intended to represent.

Overall, the field still relies heavily on task- and model-specific tests and lacks standardized

benchmarks that support direct comparison across studies. The concept of algorithmic fidelity has consequently shifted from a simple question of whether an LLM can replace human respondents to a context-sensitive, multidimensional assessment of response rates, distributions, subgroup patterns, and research goals [5]. Model selection should likewise depend on whether a study prioritizes viewpoint expression, structured output, or accurate reproduction of human attitude distributions [6]. This evolution supports a more conditional and cautious framework for evaluating LLM-based simulation.

5. Conclusion and Outlook

5.1. Research Summary

This review conceptualizes LLM-based simulation of human samples as a methodological innovation grounded in conditional generation, characterized by anthropomorphism, scalability, and controllability, and aimed at supplementing or partially replacing human participation. Its principal applications cluster in four domains: psychometrics and machine psychology, consumer and market research, experimental simulation and digital twins, and political opinion and social sentiment. Across these domains, validity is best understood through psychometric performance, group fidelity and context dependence, and ethical and epistemic justice. Figure 1 integrates these elements into the research framework synthesized in this paper.

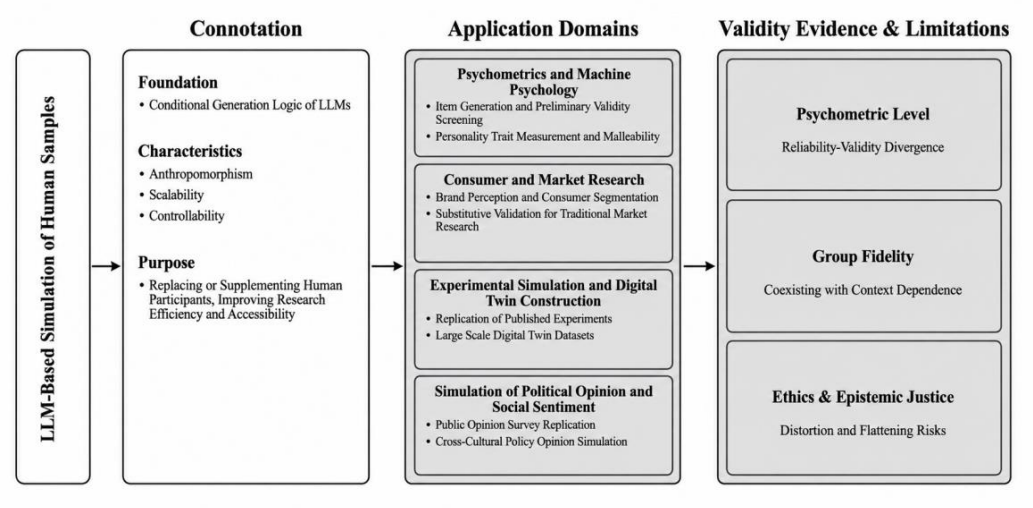

Figure 1: Research Framework for LLM-Based Simulation of Human Samples.

5.2. Outlook for Future Research

Although LLM-based simulation of human samples has produced promising results, the field remains exploratory. Future work should focus on the following issues.

First, a cross-disciplinary and standardized validity-assessment system is needed. Current studies use inconsistent indicators, tasks, and reporting practices, which limits comparability [7]. Shared benchmark sets should cover structured and open-ended tasks as well as consensual and contentious issues, while reporting model version, prompt templates, and key parameter settings to improve reproducibility [6].

Second, research should address the tension between alignment and algorithmic fidelity. Safety-oriented alignment may suppress or reshape attitude distributions that researchers seek to measure [6]. Future studies should compare base, open-source, and aligned models and evaluate whether

prompt engineering or multi-model ensembles can reduce distributional distortion while maintaining appropriate safeguards.

Third, the field needs clearer task-specific boundaries. Performance differs between structured and open-ended tasks and between low-sensitivity and culturally embedded or contentious issues [5]. Systematic comparison across task types should therefore establish when LLM simulation is sufficiently reliable for exploratory or supplementary use and when human samples remain indispensable.

Fourth, cross-cultural and cross-context validation should be strengthened, particularly in non-Western settings. Because much current evidence comes from Western contexts and systematic deviations have been observed elsewhere [5], future research should expand localized replication and evaluate calibration strategies based on local corpora and context-sensitive prompts.

Fifth, ethical and identity-justice concerns should be incorporated into methodological standards. Because LLMs may flatten or misrepresent identity groups [4], researchers should define scenarios in which synthetic substitution is inappropriate, especially when lived experience is central. Transparent reporting of model version, prompt design, alignment status, and other consequential settings is also necessary so that readers can evaluate the scope and reliability of findings.

References

[1] Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. Political Analysis, 32, 401–416. https://doi.org/10.1017/pan.2024.5

[2] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. Political Analysis, 31(3), 337–351. https://doi.org/10.1017/pan.2023.2

[3] Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., & Chen, H. (2025). Database report: Twin-2K-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. Marketing Science. https://doi.org/10.1287/mksc.2025.0262

[4] Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. Nature Machine Intelligence, 7(3), 400–411. https://doi.org/10.1038/s42256-025-00986-z

[5] Che, S., Zhu, M., Zhang, S., Jung, H. S., Lee, H., Wang, Z., & Miller, L. (2026). Simulating the people's voice: Leveraging algorithmic fidelity to assess ChatGPT's performance in modeling public opinion on Chinese government policies. Information Processing and Management, 63(1), 104567. https://doi.org/10.1016/j.ipm.2025.104567

[6] Lyman, A., Hepner, B., Argyle, L. P., Busby, E. C., Gubler, J. R., & Wingate, D. (2025). Balancing large language model alignment and algorithmic fidelity in social science research. Sociological Methods & Research, 54(3), 1110–1155. https://doi.org/10.1177/00491241251342008

[7] Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. Psychology & Marketing, 41(6), 1254–1270. https://doi.org/10.1002/mar.21982

[8] Kozlowski, A. C., & Evans, J. (2025). Simulating subjects: The promise and peril of artificial intelligence stand-ins for social agents and interactions. Sociological Methods & Research, 54(3), 1017–1073. https://doi.org/10.1177/00491241251337316

[9] Li, C., & Qi, Y. (2025). Toward accurate psychological simulations: Investigating LLMs' responses to personality and cultural variables. Computers in Human Behavior, 170, 108687. https://doi.org/10.1016/j.chb.2025.108687

[10] Zhang, J., Liang, X., Deng, A., Bonge, N., Tan, L., Zhang, L., & Zarrett, N. (2025). Leveraging interview-informed LLMs to model survey responses: Comparative insights from AI-generated and human data. arXiv preprint.

[11] Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2025). A psychometric framework for evaluating and shaping personality traits in large language models. Nature Machine Intelligence, 7, 1954–1968. https://doi.org/10.1038/s42256-025-01115-6

[12] Ferreira, G., Amidei, J., Nieto, R., & Kaltenbrunner, A. (2025). How well do simulated population samples with GPT-4 align with real ones? The case of the Eysenck Personality Questionnaire Revised-Abbreviated personality test. Health Data Science, 5, Article 0284. https://doi.org/10.34133/hds.0284

[13] Lehr, S. A., Saichandran, K. S., Harmon-Jones, E., Vitali, N., & Banaji, M. R. (2025). Kernels of selfhood: GPT-4o shows humanlike patterns of cognitive dissonance moderated by free choice. Proceedings of the National Academy of Sciences, 122(20), e2501823122. https://doi.org/10.1073/pnas.2501823122

[14] Arora, N., Chakraborty, I., & Nishimura, Y. (2025). AI–Human hybrids for marketing research: Leveraging large language models (LLMs) as collaborators. Journal of Marketing, 89(2), 1–28. https://doi.org/10.1177/00222429241276529
[15] Lim, S., Song, W., Lee, E.-J., & Jo, Y. (2025). Psychometric item validation using virtual respondents with trait-response mediators. arXiv preprint arXiv:2507.05890.
[16] Vogelsmeier, L. V. D. E., Oliveira, E., Misiejuk, K., López-Pernas, S., & Saqr, M. (2025). Delving into the psychology of machines: Exploring the structure of self-regulated learning via LLM-generated survey responses. Computers in Human Behavior, 173, Article 108769.
[17] Estevez, M., Ballestar, M. T., & Sainz, J. (2025). Market research and knowledge using generative AI: The power of large language models. Journal of Innovation & Knowledge, 10, 100796. https://doi.org/10.1016/j.jik.2025.100796
[18] Li, P., Castelo, N., Katona, Z., & Sarvary, M. (2024). Frontiers: Determining the validity of large language models for automated perceptual analysis. Marketing Science, 43(2), 254–266. https://doi.org/10.1287/mksc.2023.0454
[19] Li, Y., Liu, Y., & Yu, M. (2025). Consumer segmentation with large language models. Journal of Retailing and Consumer Services, 82, 104078.
[20] Goli, A., & Singh, A. (2024). Frontiers: Can large language models capture human preferences? Marketing Science, Articles in Advance, 1–14. https://doi.org/10.1287/mksc.2023.0306
[21] Imschloss, M., Sarstedt, M., Adler, S. J., & Cheah, J. H. (2025). Using LLMs in sensory service research: Initial insights and perspectives. The Service Industries Journal. https://doi.org/10.1080/02642069.2025.2479723
[22] Bickley, S. J., Chan, H. F., Dao, B., Torgler, B., Tran, S., & Zimbatu, A. (2025). Comparing human and synthetic data in service research: Using augmented language models to study service failures and recoveries. Journal of Services Marketing, 39(1), 36–52.
[23] Xiong, X., Wong, I. A., Huang, G. I., & Peng, Y. (2024). Understanding AI-generated experiments in tourism: Replications using GPT simulations. Journal of Travel Research, 1–17. https://doi.org/10.1177/00472875241275945
[24] Tung, Y.-H., Yang, Z.-R., Shen, M.-W., Chang, C.-Y., Chen, C.-C., & Ho, L.-C. (2025). Research note: Assessing human preferences for natural landscapes—An analysis of ChatGPT-4 and LLaVA models. Landscape and Urban Planning, 259, 105371. https://doi.org/10.1016/j.landurbplan.2025.105371
[25] Ye, Z., Yoganarasimhan, H., & Zheng, Y. (2025). LOLA: LLM-assisted online learning algorithm for content experiments. Marketing Science, Articles in Advance, 1–22. https://doi.org/10.1287/mksc.2024.0990