



Your Next Respondent Might Be an LLM

*Guidelines for
Using Silicon
Samples in
Marketing
Research*



Midjourney AI prompt by GCO



AUTHORS

Marko Sarstedt

Professor of Marketing
Institute for Marketing
Ludwig-Maximilians-University Munich
and Faculty of Economics and Business
Administration
Babeş-Bolyai University
Cluj-Napoca, Romania

Susanne J. Adler

Postdoctoral Researcher
Institute for Marketing
Ludwig-Maximilians-University Munich

Lea Rau

Research Assistant
Institute for Marketing
Ludwig-Maximilians-University Munich

Bernd Schmitt

Professor of International Business
Columbia Business School
Columbia University, New York

KEYWORDS

Generative AI
Large Language Models
Silicon Samples

Picture a market research panel that never runs late, never drops out midway and can instantly expand to thousands of participants at the click of a button. This is no futuristic fantasy; it is the emerging reality of “silicon samples,” synthetic respondents generated by large language models (LLMs). LLMs, as a form of generative artificial intelligence (GenAI), can do more than produce human-like text: They are increasingly applied to mimic human responses in interviews, focus groups or experimental scenarios. For marketing researchers, this raises an intriguing opportunity: Instead of depending solely on human participants, what if data could be generated on demand by GenAI, at scale and at low cost? Of course, with such potential come critical questions. How closely do silicon samples resemble real human behavior? In which research contexts can they be trusted, and where do they fall short? We address these questions by outlining practical use cases, suggesting a workflow with guidelines and describing key limitations to consider. ►

Key questions for generating silicon samples

FIGURE 1



GettyImages/KALA STUDIO

Silicon samples: what they are and how they are created

Silicon samples consist of synthetically generated participants that seek to mimic human respondents to describe, explain and predict human behavior, typically tailored to predefined target groups and demographic segments. Generating silicon samples is not an easy task. It requires multiple steps, including LLM selection and context adjustment, prompting techniques, sampling procedures and validation steps (Figure 1).

Use cases and potentials of silicon sampling

With their ability to simulate human responses, silicon samples can be helpful throughout the entire marketing research process and can be readily integrated into standard workflows. Academic research on this topic is quickly evolving, and here are the applications current findings suggest:

Pretesting stimuli and survey items

LLMs can function as advisors in pretesting and pilot studies. They can offer alternative wordings and help refine research materials, for example, by identifying double-barreled questions and vague

response options. Multimodal LLMs can also provide feedback on visual stimuli by assessing visual properties, aligning visual elements with intended meanings and suggesting improvements. Prompting can further stimulate various cultural and personal perspectives and help ensure similar perceptions across target segments.

Generating large-scale data for quantitative research

Silicon samples offer the potential to supplement or replace large-scale quantitative data. LLMs could leverage up-to-date input data, such as from social media or review platforms, for continuous benchmarking, allowing for real-time monitoring of, for example, customer mindset metrics. With their fast and scalable data generation, LLMs can also help restock data collections and extend existing samples.

Generating synthetic personas for qualitative insights

LLMs may be used to design synthetic personas that respond to qualitative interviews, allowing researchers to collect data from prolonged interviews, explore alternative conversation paths or ask additional questions at a later stage. Such qual-

itative inquiries can be extended to focus groups, where synthetic personas discuss topics under the guidance of a human moderator.

Challenges and limitations of silicon sampling

Despite these advantages, using silicon samples remains highly challenging, which limits their usefulness for strategic decision-making. Standard LLMs, in particular, are currently not designed to replicate the individual lived experiences, cultural nuances and behaviors that characterize genuine human decision-making.

Domain-specific (under)performance

In our recent literature review across 285 silicon-to-human sample comparisons, we found that 24.9% of comparisons yielded similar results, while 65.3% diverged and 9.8% only partially aligned. Although LLMs could replicate outcomes related to some stable constructs, such as personality traits and political preferences, their output often did not align with well-established consumer behavior effects.

Skewed distributions from training data

The output quality is inherently dependent on the input training data – a manifestation of the classic “garbage in, garbage out” principle. LLMs are commonly trained on open-source, English-language textual data from platforms such as Wikipedia and GitHub. Their responses therefore tend to reflect the perspectives and characteristics of Western societies, making them less reliable for research focused on other populations.

Sensitivity to prompt design

The behavior of LLMs is sensitive to prompt design. Subtle variations in prompt structure, such as response order, labeling and framing, can affect model output and introduce systematic biases.

Missing variance

LLM responses frequently lack variance. Many research projects, however, exploit variance to uncover relationships and segments. Choosing an appropriate prompting technique can infuse variability into a silicon sample, but this information is deliberately chosen and neglects unforeseen effects that would emerge in a human sample.

Guidelines for generating valid silicon samples in marketing research

Given the previously outlined limitations of silicon samples, users should not take an LLM’s output at face value when making decisions. Instead, producing valid silicon samples requires careful scrutiny, including researchers’ critical evaluation of LLM outputs and benchmarking against human-generated data at both the level of individual synthetic respondents and the overall silicon sample (Figure 2). Model selection and prompt design require special attention. LLMs vary in their ability to replicate human response patterns. LLMs that are adjusted to user-specific requirements generally outperform out-of-the-box models. We therefore recom-

Standard LLMs are currently not designed to replicate the individual lived experiences that characterize genuine human decision-making.



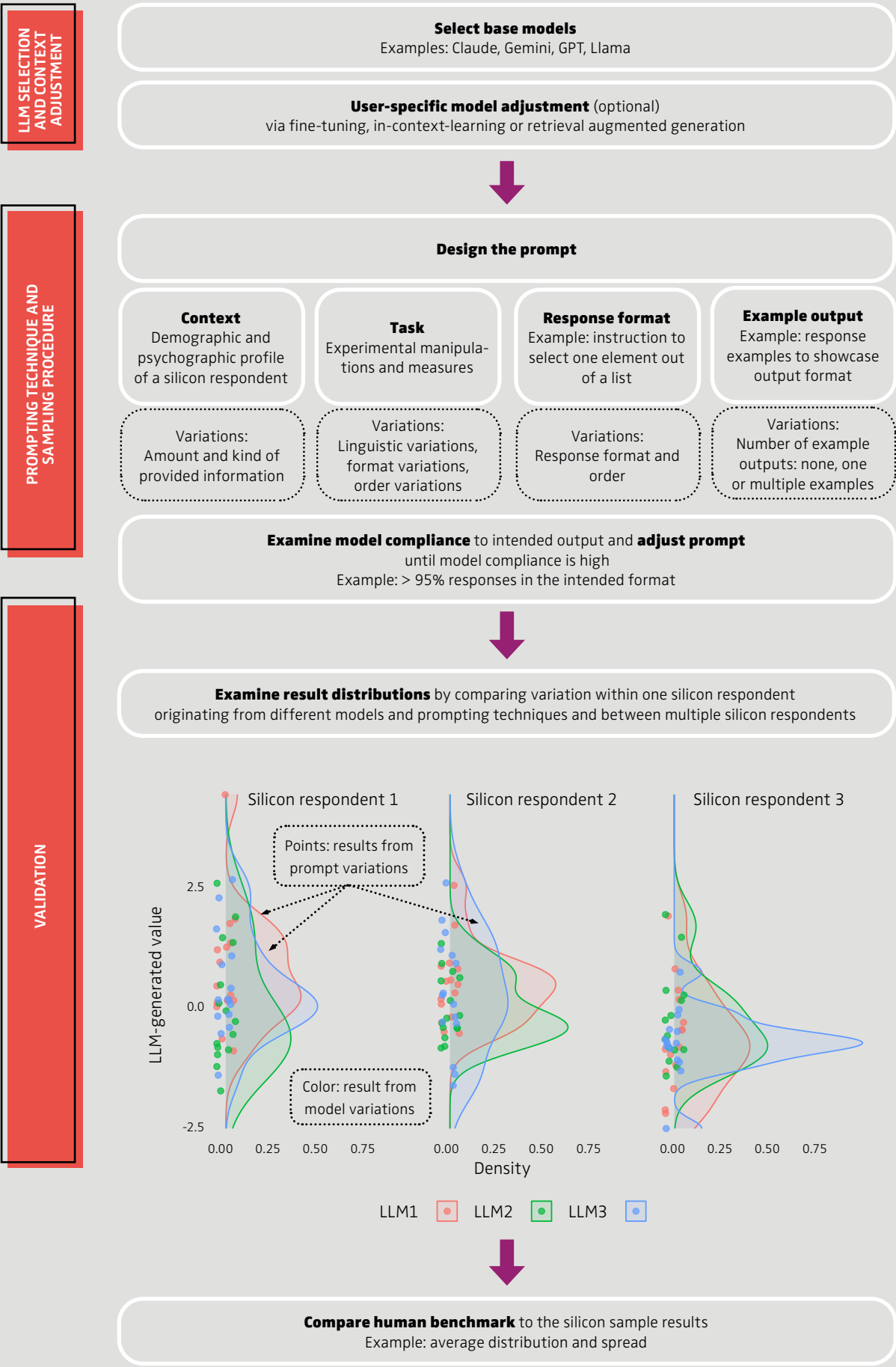
Midjourney AI prompt by GCO

mend the use of multiple models and, preferably, the induction of additional context information via fine-tuning or retrieval-augmented generation (RAG). Similarly, prompt design is crucial to a silicon sample’s validity. A prompt should include information on the context, the task, as well as the response format and should be optimized until the LLM delivers the intended output format. Results from varying prompts should be examined according to their distribution (see Figure 2). What is the distribution’s “average” result, such as its mean, median or most common value? How variable are the results in terms of their variance or range? This distribution can then be compared against a human benchmark distribution. The closer the LLM-generated distribution is to the human values, the better the silicon sample.

Our discussion so far has shown that silicon samples hold great potential, but their generation and use are not straightforward processes. Figure 3 summarizes our discussion in the form of guidelines. ►

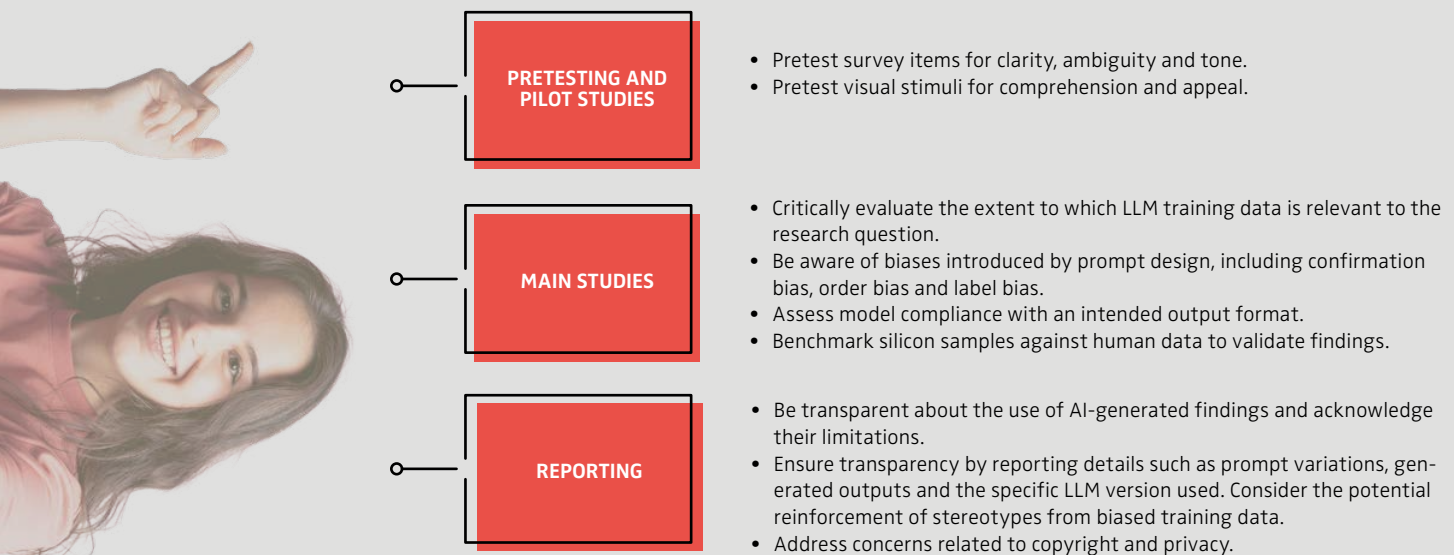
Workflow to design valid silicon samples

FIGURE 2



Guidelines for research projects using silicon sampling

FIGURE 3



GettyImages/KALA STUDIO

The future lies in specialized models for specific tasks

LLMs may better approximate human behavior soon, especially once they integrate multimodal inputs such as image, video and voice and develop more nuanced human-like reasoning capabilities – a topic that has recently been increasingly debated in the academic literature. Advances in prompt design and context infusion to LLMs will also contribute to advancing silicon samples as human proxies. Unlike with human data, researchers cannot assume that an LLM's responses will automatically consider a lifetime of experiences and personal development. We do not know exactly how the mind works, nor how similar LLMs are to the human mind. What is clear, however, is that LLM architectures were not built in the same way the human mind evolved through biological processes. While LLMs can mimic human responses, they do not have a nervous system and thus may “experience” the world differently, thus perhaps lacking the very foundation for human judgment. This divergence might fundamentally limit an LLM's ability to capture human decision-making, emotional responses and cultural diversity. For this reason, validation is likely to remain crucial when working with silicon samples.

The premier, one-size-fits-all models like the latest versions of GPT, Llama or Claude will also likely continue to struggle with specialized tasks such as silicon sampling. Substantial advances – such as the leap from GPT-3.5 to GPT-4o – would be necessary to enhance silicon sampling capabilities, but such progress seems unlikely in upcoming models, as suggested by the partly disappointed reactions to GPT-5. The era of exponential growth in LLM performance may be over, at least for now. Instead, fragmentation into specialized LLMs that are fine-

tuned to excel at a limited set of tasks based on context-specific data seems more likely. In terms of behavioral research, the Centaur model, which is a fine-tuned Llama model, for example, surpassed its base model in terms of human response prediction. Resource-efficient approaches to in-context learning or RAG further make model adjustments affordable and accessible to a large customer base. We therefore expect to see more tailored models that perform extraordinarily on one specific task, such as mimicking responses from a specific target group. Early versions of such personalized models already exist, and training LLMs on a large amount of personal data will likely result in even better silicon samples. Reaping this potential requires close collaborations between marketing researchers, psychologists and computer scientists to ensure that insights remain relevant to real human behavior.

FURTHER READING

Brucks, M., & Toubia, O. (2025).

Prompt architecture induces methodological artifacts in large language models. *PLoS One*, 20(4), <https://doi.org/10.1371/journal.pone.0319159>

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2025).

Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24), <https://doi.org/10.1073/pnas.2501660122>

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 and Marketing*, 41(6), 1254–1270. <https://doi.org/10.1002/mar.21982>

Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., & Chen, H. (2025).

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*, 44(6), 1446–1455. <https://doi.org/10.1287/mksc.2025.0262>