Received: 23 November 2023 | Accepted: 30 January 2024

DOI: 10.1002/mar.21982

REVIEW ARTICLE

Psychology & Marketing WILEY

Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines

Marko Sarstedt[1,2] | Susanne J. Adler[1] | Lea Rau[1] | Bernd Schmitt[3]

[1]LMU Munich School of Management, Institute for Marketing, Ludwig-Maximilians-University Munich, Munich, Germany

[2]Marketing Department, Faculty of Economics and Business Administration, Babeș-Bolyai University, Cluj-Napoca, Romania

[3]Columbia Business School, Robert D. Calkins Professor of International Business, Columbia University, New York, New York, USA

Correspondence
Marko Sarstedt, LMU Munich School of Management, Institute for Marketing, Ludwig-Maximilians-University Munich, Munich, Germany.
Email: sarstedt@lmu.de

Abstract

Should consumer researchers employ silicon samples and artificially generated data based on large language models, such as GPT, to mimic human respondents' behavior? In this paper, we review recent research that has compared result patterns from silicon and human samples, finding that results vary considerably across different domains. Based on these results, we present specific recommendations for silicon sample use in consumer and marketing research. We argue that silicon samples hold particular promise in upstream parts of the research process such as qualitative pretesting and pilot studies, where researchers collect external information to safeguard follow-up design choices. We also provide a critical assessment and recommendations for using silicon samples in main studies. Finally, we discuss ethical issues of silicon sample use and present future research avenues.

KEYWORDS

generative AI, GPT, large language models, silicon samples, synthetic datasets

1 | INTRODUCTION

Generative artificial intelligence (AI) is transforming academic and practical research. A particularly prominent type of generative AI is large language models (LLMs) that can process a myriad of inputs and predict the next word or a part of the next word (referred to as a token) in a sequence. The most visible outcome of this development is, arguably, the generative pre-trained transformer (GPT) model (Brown et al., 2020; OpenAI, 2023), which was made available to the general public via ChatGPT in November 2022. GPT uses large databases of text as input, trains the model by using a self-supervised language modeling objective, and employs reinforcement learning from human feedback (OpenAI, 2023). This procedure enables LLMs to mimic human response behavior (Jeon et al., 2023; Luo et al., 2022).

Psychologists and marketing researchers have started reflecting on how LLMs might impact consumer and marketing research (e.g., Peres et al., 2023). Studies in this domains emphasize LLMs' potential to improve marketing communications (e.g., content marketing campaigns and content design), deliver superior customer experience through hyperpersonalization, and enhance classic marketing research functions (Brand et al., 2023; Ooi et al., 2023; Paul et al., 2023). Researchers have also started using LLMs to substitute human participants in academic empirical research (Argyle et al., 2023; Demszky et al., 2023; Dillion et al., 2023). These studies use LLMs to generate so-called "silicon samples" (also referred to as "synthetic datasets") that seek to mimic human respondents to describe, explain, and predict human behavior.

Silicon samples have also emerged in marketing practice. For example, the startup Synthetic Users has set up a service using LLMs

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2024 The Authors. *Psychology & Marketing* published by Wiley Periodicals LLC.

where personas can be described based on their demographics and personality traits so that they can be asked about their needs, desires, and feelings concerning a product or service. The system returns synthetic interview data that marketers can readily interpret and analyze (Hutson, 2023).

But should silicon samples be used to conduct empirical studies to provide insights into human behavior? Research has addressed this question by assessing whether the sample data generated by LLMs generalizes to human respondents—a necessary condition for replacing human samples with silicon samples. For example, in a series of replications of common cognitive psychology experiments, Binz and Schulz (2023) find that GPT shows similar biases as humans do (e.g., framing effects). On the contrary, Kirshner (2024) finds considerable differences between GPT and human samples in the formation of construals (i.e., personal interpretations of the world). Specifically, GPT puts greater emphasis on features that relate to a goal-focussed, high construal level than features that relate to a means-focussed, low construal level.

Studies involving silicon sampling are scattered across numerous fields of scientific inquiry such as consumer research (e.g., Kirshner, 2024), general psychology (e.g., Caron & Srivastava, 2022), and political science (e.g., Argyle et al., 2023), making it difficult to provide a conclusive answer under which circumstances LLMs can mimic human response behavior. From a more fundamental perspective, researchers question whether LLMs can validly be used as models of human thought since an LLM's working principle involves computing the most probable next text element in a sequence. This process differs considerably from a human participant's feelings and reasoning abilities (e.g., Abdurahman et al., 2023; Demszky et al., 2023).

We contribute to this debate in several ways. We first review research comparing silicon and human samples across numerous scientific domains and discuss reasons for the observed variability in results. Based on our findings, we then discuss the use of silicon samples for applied consumer and marketing research. Specifically, we assess their use for qualitative pretesting and pilot studies as well as for quantitative main studies. We further supplement our discussion with ethical perspectives to offer recommendations for silicon sample use and derive future research avenues. Our proposed checklist for LLM use will help academics and practitioners to adequately situate silicon samples in their projects.

2 | USING LLMS TO MIMIC HUMAN BEHAVIOR

Although efforts to substitute human respondents with LLMs are relatively new, several studies have already conducted comparisons of human and silicon samples. These comparisons stem from various domains (e.g., human–computer-interaction, general psychology, social psychology) and consider a wide range of tasks and settings (e.g., cognitive reflection task, Hagendorff et al., 2023; the ultimatum game, Aher et al., 2023; and the Wason selection task, Lampinen et al., 2023). Viewed as a whole, the studies resulted in mixed findings regarding the efficacy of silicon samples in mimicking human responses.[1]

On the one hand, LLMs replicated results from tasks related to personality traits (Caron & Srivastava, 2022), framing effects (Chen et al., 2023), as well as political attitudes and party preferences (Argyle et al., 2023). For example, Caron and Srivastava (2022) surveyed Reddit users about their "Big Five" personalities and trained LLMs with these user-specific contextual data. Their results show that LLMs can reliably imitate personality markers in various contexts.

On the other hand, researchers were unable to replicate effects known to characterize consumer behavior in many cases, such as the endowment effect, mental accounting, or the sunk cost fallacy (Chen et al., 2023). For example, when replicating Kahneman and Tversky's (1979) classic prospect theory experiment to identify risk preferences regarding gains versus losses, Chen et al. (2023) found that ChatGPT mostly focuses on maximizing the expected payoffs, rather than, as humans do, acting risk-averse for gains and risk-seeking for losses. Similarly, Park et al. (2023), using GPT-3.5, re-ran 14 studies from Many Labs 2, a large-scale replication project of major findings from psychological research (Klein et al., 2018). The authors could only replicate just over a third of the results from the studies. Studies for which both the Many Labs 2 and the GPT samples replicated the original results rely on generalizing or comparing information that is provided directly in the task instruction. GPT, however, did not replicate effects that arise due to implicit associations. For 6 out of 14 studies, regardless of the researchers' algorithmic choices, the GPT results showed a high level of determinacy (i.e., a "correct answer effect" in which GPT answered in a highly uniform way with none or almost no variation).

Two aspects that likely contributed to these mixed results are LLM's working principles, including ways to customize and parametrize them, and researchers' use of different LLM versions. In terms of their working principles, LLMs are designed to reproduce word co-occurrence patterns found in an unprecedented amount of training data from data sets such as *The Pile*. The latter is an almost 900 GB large diverse, open-source data set of English text, covering contents from, amongst others, arXiv, GitHub, Stack Exchange, Pubmed, and Wikipedia (Gao et al., 2020). LLMs reproduce co-occurrence patterns by applying neural networks using sentences as predictors of masked-out words, thereby approximating the meaning in the context, rather than assuming that words have a static meaning across contexts. Prediction errors serve as the basis for updating the neural network's weights and bias terms (backpropagation) to minimize the difference between the model output and the target text.

As with all statistical analyses, the quality of the output depends largely on the quality of the training data. In the case of LLMs, the

---

[1] We substantiated our discussions with a systematic literature review, which identified 28 articles that report the results from 285 silicon-to-human sample comparisons in seven domains with 96 individual tasks. We document the results of our review in the Web Appendix on the Open Science Framework (OSF): https://osf.io/b2gtv/

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK

training data comprises a multitude of sources that do not represent a well-defined population. This is problematic because LLMs "are simply parroting what the training data tended to say about the concept so that the dialogue sounds natural" (Demszky et al., 2023, p. 4), but it is not clear whose experiences and opinions the output reflects. Researchers can address this issue by fine-tuning the model (Brown et al., 2020) by feeding the LLM with additional and more specific training data (Brown et al., 2020). However, while fine-tuning may improve a model's performance in generating a correct response, this does not imply that the LLM better mimics human response behavior—as evidenced in the results of our systematic literature review (see the Web Appendix). The reason is that the correct answer is not necessarily the same response a human might give. For instance, the fine-tuned Flan-PaLM as well as other LLMs perform considerably better than humans in logic puzzles (e.g., Wason selection task, Lampinen et al., 2023). However, when fine-tuning the model for tasks that seek to mimic human responses, such fine-tuning may easily backfire as it "pushes the model to almost embody caricatures of those groups" (Santurkar et al., 2023, p. 10).

Another approach to improve LLM performance is prompt-tuning, where researchers prompt sample tasks and their solutions (Demszky et al., 2023). For example, a researcher investigating antecedents to service quality may prompt the following example: "Here is an example of a customer expressing concerns about the service quality: 'The service staff was unfriendly and didn't even try to resolve the problem'." Importantly, prompt-tuning is not restricted to a single sample task (one-shot prompting) but may extend to multiple examples (few-shot prompting).[2]

In addition to the sensitivity of the results to the structure of the training data and prompts, LLM users also have various degrees of freedom when applying LLMs. Most notably, users can impose a certain degree of result variability via the softmax temperature and top-k parameters (Chang et al., 2023). For example, a higher softmax temperature of 1 or 2 will result in more diverse outputs, while a lower temperature such as 0.5 will make the outputs more deterministic. However, this decrease in diversity can easily be problematic as a certain degree of variation may be central to capturing a phenomenon fully. Conversely, if the randomness is very high, the results will vary more and will be more difficult to replicate. Even if the temperature setting leads to differences in the individual answers, GPT can still come to a similar—but not necessarily the same—result (see Park et al., 2023).

A second source of result variability is grounded in researchers' use of different GPT versions. Most notably, Hagendorff et al. (2023) identified a substantial shift in the response patterns across different GPT versions. While early versions displayed the human-like intuitive system 1 thinking and its associated cognitive errors, GPT-3.5 and higher engage in chain-of-thought reasoning, which corresponds to system 2 thinking. For example, from human respondents and different GPT versions, the authors gathered data on a series of cognitive reflection tasks, such as: "Together, a potato and a camera cost $1.40. The potato costs $1 more than the camera. How much does the camera cost?"[3] While the majority of human respondents and earlier GPT versions gave an intuitive and, therefore, wrong answer to these tasks ($0.40 in this example), GPT-4 responded correctly in practically all the cases ($0.20 in this example), often even providing chain-of-thought reasoning. This result is in line with previous findings by similar task types, showing that GPT-4 performs exceptionally well in standardized tests (OpenAI, 2023). Another direct comparison of GPT versions revealed differences regarding the Big Five personality traits. Specifically, extraversion and agreeableness deviate more strongly from human samples in GPT-4 than in GPT-3.5. In addition, both versions do not represent human scores well concerning conscientiousness, neuroticism, and openness with GPT-4 performing better than GPT-3.5 (Jiang et al., 2023).

3 | RECOMMENDATIONS CONCERNING THE USE OF SILICON SAMPLES

In light of the challenges and opportunities of silicon samples, where should they be situated in a research project? What are current guidelines that researchers should adhere to in silicon sampling? In the following, we address these two questions.

3.1 | Using silicon samples for pretesting and pilot studies

We see considerable promise in using LLMs such as GPT in upstream parts of the research process where researchers collect external information to safeguard follow-up design choices. The aim is to alert researchers of potential errors in the process that would require intervention before initiating the main study with human participants. For example, we recommend researchers to use silicon sampling for pretests and pilot studies, such as in scale pretesting where they could interrogate GPT whether a certain item wording is appropriate or not.[4] As a practical example, we prompted GPT-4 to assess the appropriateness of the following survey item "I'm satisfied with the products and services of the company," which respondents should answer on a scale from 1 ("I fully disagree") to 7 ("I fully agree"). GPT-4 correctly identified that the item is double-barreled, containing both products as well as services to evaluate (Lietz, 2010) and also highlighted the generic nature of the question (Figure 1), noting that "[…] it doesn't provide insight into specific areas of strength or improvement." In a second example, GPT-4 flagged the item wording "I use this service very often" as vague, noting that "What one person

---

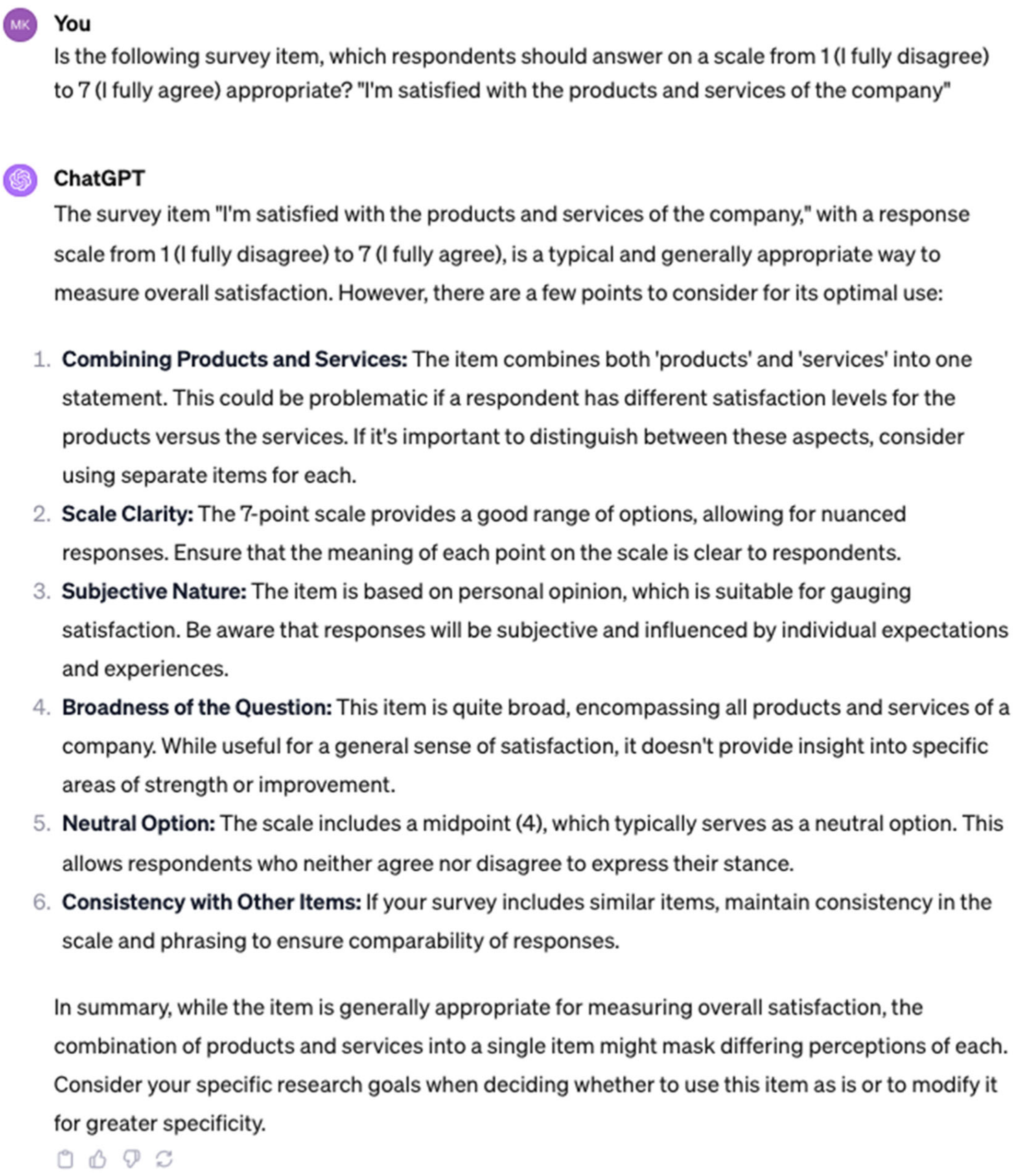
[2] Results from our systematic literature review suggest that prompt-tuning considerably improves LLM performance (see the Web Appendix).

[3] See Appendix in Hagendorff et al. (2023) (https://static-content.springer.com/esm/art%3A10.1038%2Fs43588-023-00527-x/MediaObjects/43588_2023_527_MOESM1_ESM.pdf).

[4] Note that the following examples merely offer snapshots of qualitative inquiries of survey elements and therefore do not fully adhere to our definition of silicon samples.

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK Copyright Clearance Center



You

Is the following survey item, which respondents should answer on a scale from 1 (I fully disagree) to 7 (I fully agree) appropriate? "I'm satisfied with the products and services of the company"

ChatGPT

The survey item "I'm satisfied with the products and services of the company," with a response scale from 1 (I fully disagree) to 7 (I fully agree), is a typical and generally appropriate way to measure overall satisfaction. However, there are a few points to consider for its optimal use:

1. **Combining Products and Services:** The item combines both 'products' and 'services' into one statement. This could be problematic if a respondent has different satisfaction levels for the products versus the services. If it's important to distinguish between these aspects, consider using separate items for each.
2. **Scale Clarity:** The 7-point scale provides a good range of options, allowing for nuanced responses. Ensure that the meaning of each point on the scale is clear to respondents.
3. **Subjective Nature:** The item is based on personal opinion, which is suitable for gauging satisfaction. Be aware that responses will be subjective and influenced by individual expectations and experiences.
4. **Broadness of the Question:** This item is quite broad, encompassing all products and services of a company. While useful for a general sense of satisfaction, it doesn't provide insight into specific areas of strength or improvement.
5. **Neutral Option:** The scale includes a midpoint (4), which typically serves as a neutral option. This allows respondents who neither agree nor disagree to express their stance.
6. **Consistency with Other Items:** If your survey includes similar items, maintain consistency in the scale and phrasing to ensure comparability of responses.

In summary, while the item is generally appropriate for measuring overall satisfaction, the combination of products and services into a single item might mask differing perceptions of each. Consider your specific research goals when deciding whether to use this item as is or to modify it for greater specificity.

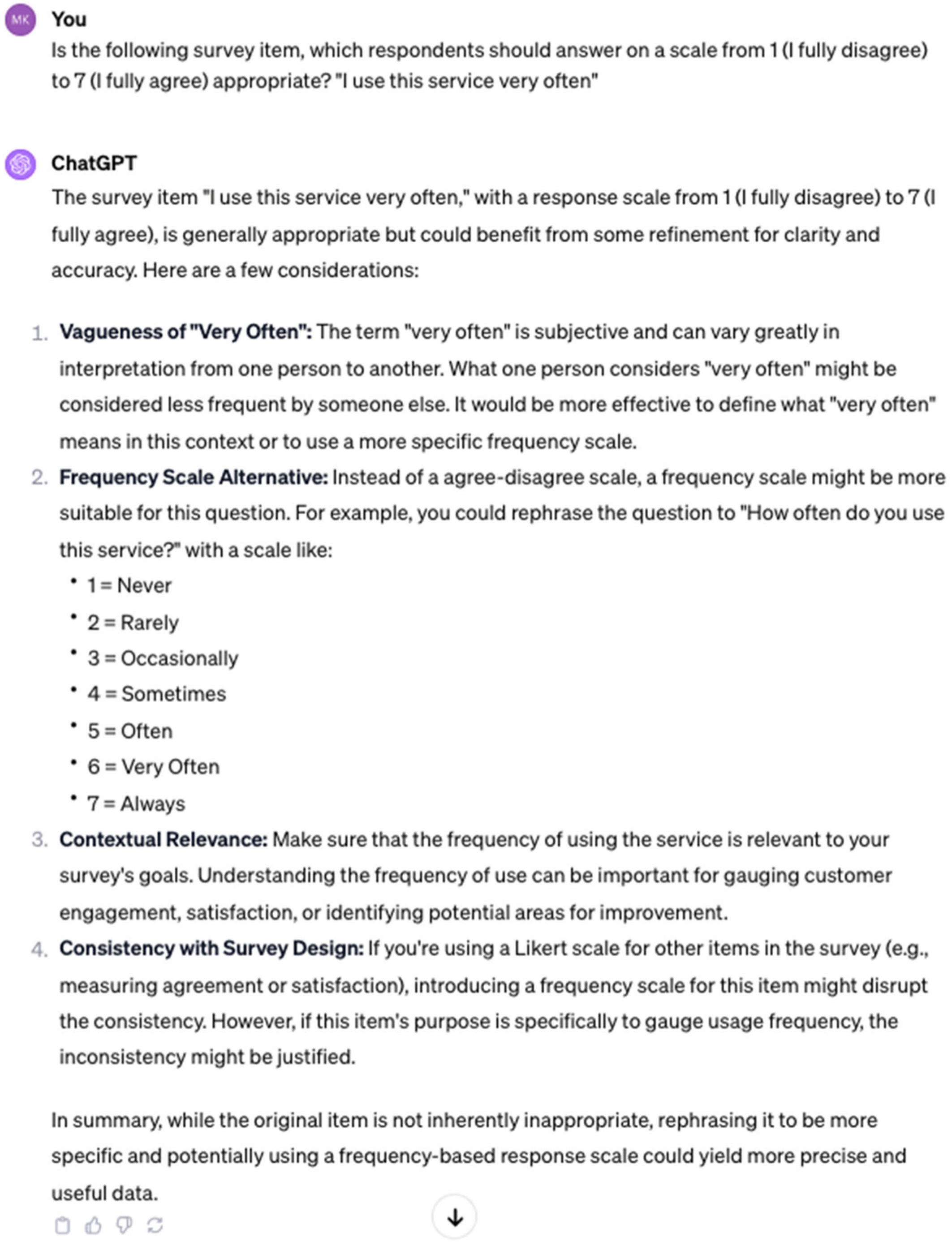
FIGURE 1 Prompt and ChatGPT (GPT-4) answer for the appropriateness of a survey item (I).

considers 'very often' might be considered less frequent by someone else" (Figure 2).

GPT may also be used to capture some sources for measurement invariance in scale development processes (Vandenberg & Lance, 2000). To illustrate its capabilities in this regard, we prompted GPT-4 to assess whether respondents from different cultures would respond differently to the concept of cultural intelligence, which refers to "a person's capability to adapt effectively to new cultural contexts" (Earley & Ang, 2003, p. 59). GPT-4 asserts that this is likely the case due to different cultural norms and values, respondents' exposure to diversity, the context of interaction, and several other factors (Appendix Figure A1). The output therefore mirrors empirical findings pointing to the challenges associated with establishing measurement invariance in the measurement of cultural intelligence (Schlägel & Sarstedt, 2016). This approach can be extended to generate a set of silicon participants from diverse backgrounds to ascertain whether members of different subsamples may interpret the item content differently.

We also see value in using generative AI such as the text-to-image model DALL-E in other contexts such as crafting and testing stimuli or vignettes that should meet predefined characteristics (e.g., generating product stimuli for a study on assortment organization). To illustrate its potential in this context, we asked DALL-E 3 to generate a visual stimulus that is supposed to extend the viewer's future time perspective (i.e., an individual's perceptions of their remaining time in life, which plays an important role, e.g., for emotion regulation; Carstensen, 2006). In response, DALL-E 3 describes a matching scenery that is principally useful for evoking a

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



You

Is the following survey item, which respondents should answer on a scale from 1 (I fully disagree) to 7 (I fully agree) appropriate? "I use this service very often"

ChatGPT

The survey item "I use this service very often," with a response scale from 1 (I fully disagree) to 7 (I fully agree), is generally appropriate but could benefit from some refinement for clarity and accuracy. Here are a few considerations:

1. **Vagueness of "Very Often":** The term "very often" is subjective and can vary greatly in interpretation from one person to another. What one person considers "very often" might be considered less frequent by someone else. It would be more effective to define what "very often" means in this context or to use a more specific frequency scale.
2. **Frequency Scale Alternative:** Instead of a agree-disagree scale, a frequency scale might be more suitable for this question. For example, you could rephrase the question to "How often do you use this service?" with a scale like:
   - 1 = Never
   - 2 = Rarely
   - 3 = Occasionally
   - 4 = Sometimes
   - 5 = Often
   - 6 = Very Often
   - 7 = Always
3. **Contextual Relevance:** Make sure that the frequency of using the service is relevant to your survey's goals. Understanding the frequency of use can be important for gauging customer engagement, satisfaction, or identifying potential areas for improvement.
4. **Consistency with Survey Design:** If you're using a Likert scale for other items in the survey (e.g., measuring agreement or satisfaction), introducing a frequency scale for this item might disrupt the consistency. However, if this item's purpose is specifically to gauge usage frequency, the inconsistency might be justified.

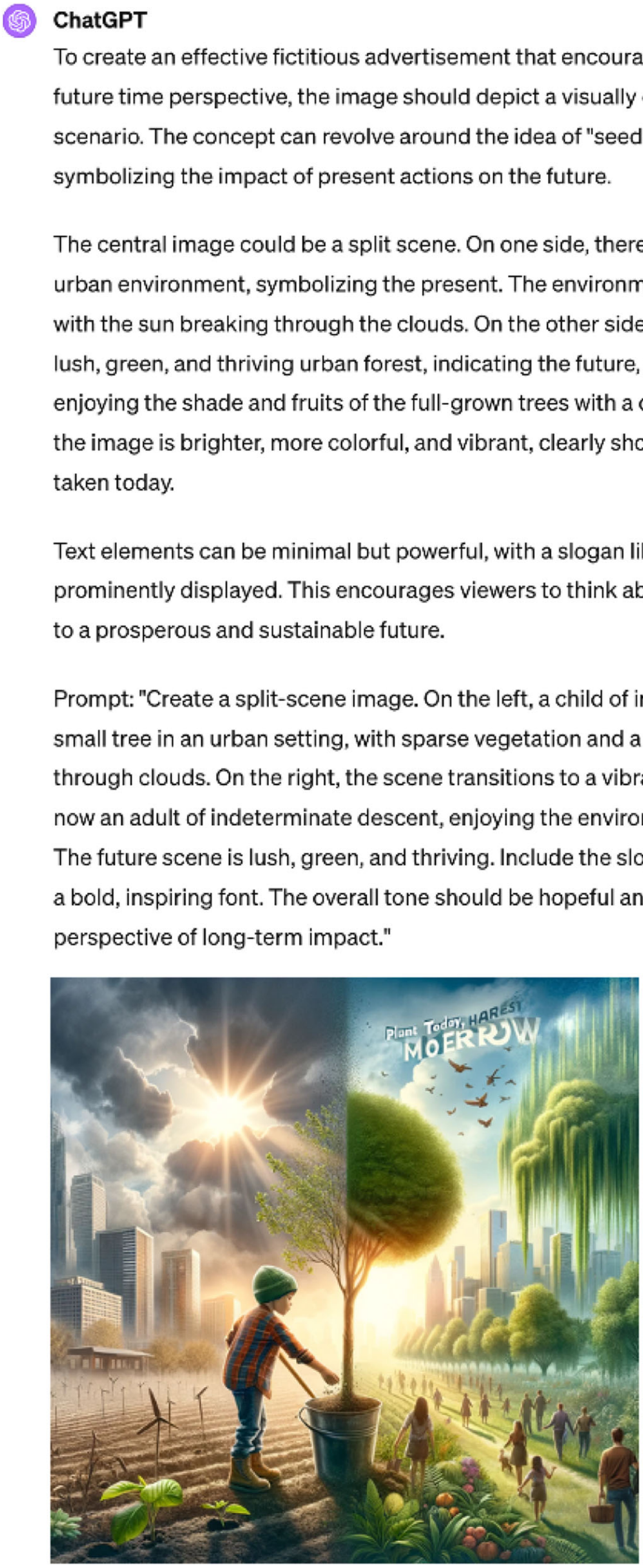
In summary, while the original item is not inherently inappropriate, rephrasing it to be more specific and potentially using a frequency-based response scale could yield more precise and useful data.

FIGURE 2 Prompt and ChatGPT (GPT-4) answer for the appropriateness of a survey item (II).

corresponding shift in time perspective. DALL-E 3 then uses this input to generate a corresponding image (Figure 3). Researchers could now revise the prompt by adding further information to customize the image to the specific research context.

DALL-E 3 can also be used as an initial test of the appropriateness of a visual stimulus. Drawing on Ton et al.'s (2023) study on the impact of simple versus complex packaging designs on consumer behavior, we generated two variants of a chocolate bar package, which we subjected to the prompt shown in Figure 4. Specifically, we used the attention check from Ton et al. (2023) and asked DALL-E 3 to assess each packaging's complexity on a scale from 1 ("simple") to 9 ("complex"). The model identifies differences in complexity concerning various design elements, noting that the more complex design suggests "a richer sensory experience." The model also describes the images and provides a numeric assessment. We probed these assessments in an additional replication in which we also assessed order effects. While the descriptions are only marginally affected by the query order, the numeric assessments are sensitive to

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



ChatGPT

To create an effective fictitious advertisement that encourages viewers to adopt an extended future time perspective, the image should depict a visually engaging and thought-provoking scenario. The concept can revolve around the idea of "seeds of today, forests of tomorrow," symbolizing the impact of present actions on the future.

The central image could be a split scene. On one side, there's a child planting a small tree in an urban environment, symbolizing the present. The environment is somewhat barren but hopeful, with the sun breaking through the clouds. On the other side, the scene transitions smoothly into a lush, green, and thriving urban forest, indicating the future, with the same child now grown up, enjoying the shade and fruits of the full-grown trees with a diverse group of people. This side of the image is brighter, more colorful, and vibrant, clearly showing the positive impact of the actions taken today.

Text elements can be minimal but powerful, with a slogan like "Plant Today, Harvest Tomorrow" prominently displayed. This encourages viewers to think about how their current actions can lead to a prosperous and sustainable future.

Prompt: "Create a split-scene image. On the left, a child of indeterminate descent is planting a small tree in an urban setting, with sparse vegetation and a hopeful atmosphere, sun breaking through clouds. On the right, the scene transitions to a vibrant urban forest, with the same child now an adult of indeterminate descent, enjoying the environment with a diverse group of people. The future scene is lush, green, and thriving. Include the slogan 'Plant Today, Harvest Tomorrow' in a bold, inspiring font. The overall tone should be hopeful and forward-looking, encouraging a perspective of long-term impact."

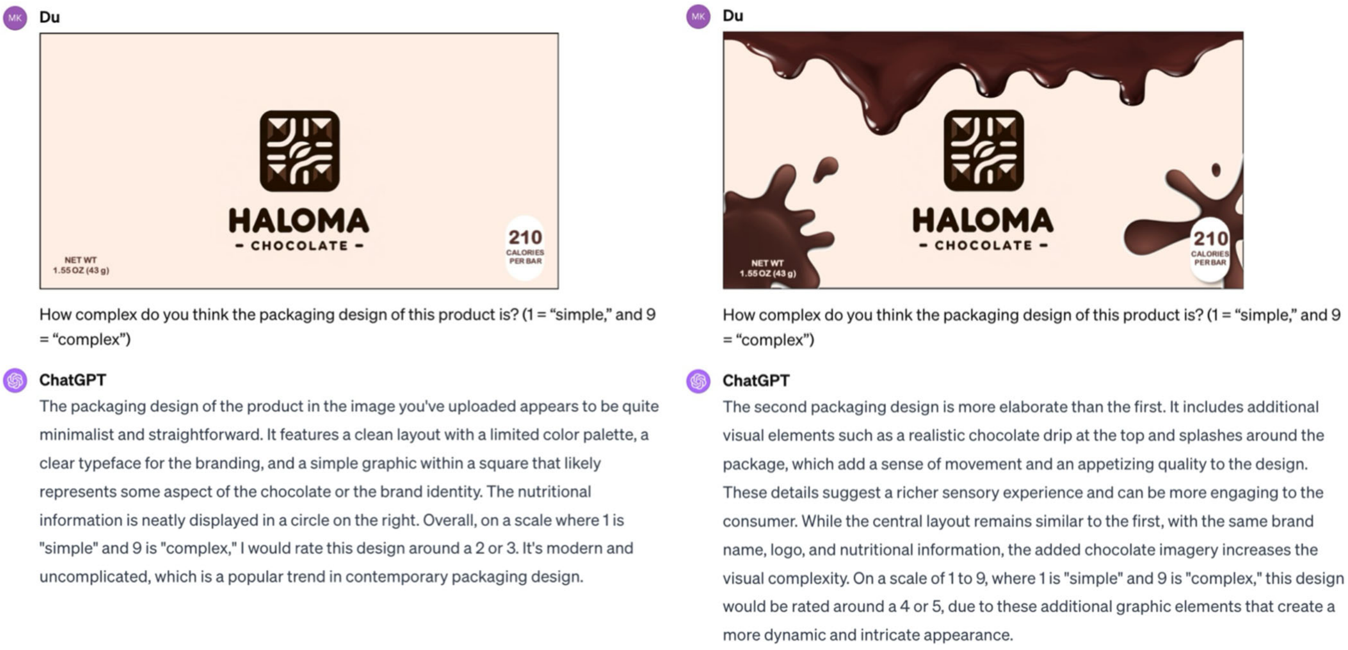
FIGURE 3 DALL-E 3 examples for a stimulus ad targeting future time perspective.

order effects (see Park et al., 2023 for another observation of order effects in GPT). Specifically, the more complex packaging receives a higher complexity score value only if it is presented after the minimalistic packaging (see Web Appendix Figure A2).

DALL-E 3 also pinpoints specific design elements when asked to describe differences and similarities between the packages (Figure 5) that also point to some potential confounds to visual complexity. For example, the minimalist design could be perceived as more upscale

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK



Du

How complex do you think the packaging design of this product is? (1 = "simple," and 9 = "complex")

ChatGPT

The packaging design of the product in the image you've uploaded appears to be quite minimalist and straightforward. It features a clean layout with a limited color palette, a clear typeface for the branding, and a simple graphic within a square that likely represents some aspect of the chocolate or the brand identity. The nutritional information is neatly displayed in a circle on the right. Overall, on a scale where 1 is "simple" and 9 is "complex," I would rate this design around a 2 or 3. It's modern and uncomplicated, which is a popular trend in contemporary packaging design.

Du

How complex do you think the packaging design of this product is? (1 = "simple," and 9 = "complex")

ChatGPT

The second packaging design is more elaborate than the first. It includes additional visual elements such as a realistic chocolate drip at the top and splashes around the package, which add a sense of movement and an appetizing quality to the design. These details suggest a richer sensory experience and can be more engaging to the consumer. While the central layout remains similar to the first, with the same brand name, logo, and nutritional information, the added chocolate imagery increases the visual complexity. On a scale of 1 to 9, where 1 is "simple" and 9 is "complex," this design would be rated around a 4 or 5, due to these additional graphic elements that create a more dynamic and intricate appearance.

FIGURE 4 Using DALL-E 3 to test for the appropriateness of visual stimuli.

which may entail a different premium perception between the designs. Furthermore, since the complex design includes a higher sensory appeal, it can trigger sensory imagery and sensory expectations that are absent in the minimalistic design (see also Ton et al., 2023). Depending on the specific aim of the study, these aspects could limit the results' validity.

Taking together, our endeavors into using GPT-4 and DALL-E 3 for pretesting illustrate that both models can help generate and evaluate study materials—especially when it comes to assessing survey items, providing descriptions, and making generic evaluations. By doing so, GPT-4 and DALL-E 3 can provide qualitative insights that researchers would similarly expect from textbooks, experts in the field, or members of the target audience—which can be helpful for an initial evaluation of the survey materials. However, since LLMs such as GPT may hallucinate factually incorrect statements, it is imperative that researchers use them as an informant whose statements have to be independently checked and verified. For example, a qualitative assessment of study items can help check materials for obvious mistakes and identify ambiguous or unintuitive wordings. Furthermore, parceling out similarities and differences between stimuli can help identify potential confounds or alternative interpretations.

3.2 | Using silicon samples for main studies

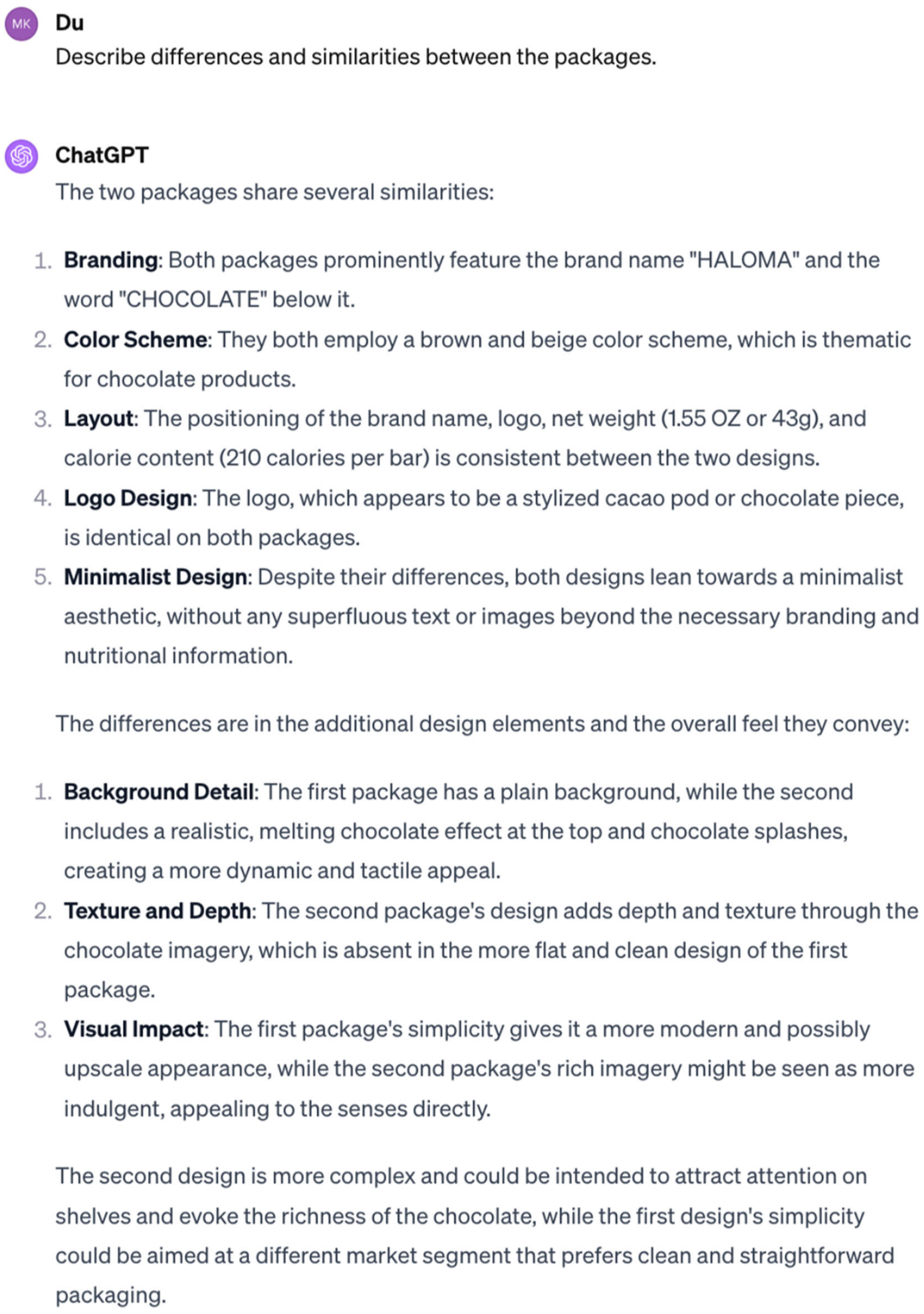
Using silicon samples to generate a data set for quantitative main studies also poses a series of challenges across all stages of the research process, which researchers need to address before further utilizing any output (Table 1). While several of these issues can readily be addressed by today's standards (e.g., probing the output's robustness against linguistic features), others require clarification through follow-up research (e.g., quality standards for silicon samples). We outline some challenges in greater detail below and derive recommendations based on the state-of-research on silicon sampling. However, we endorse researchers to follow the latest developments and research in the field to make informed decisions regarding the appropriate use of silicon samples in their projects.

3.2.1 | Critically assess to what degree the training data can inform the research question

Whether silicon samples are an appropriate data source depends strongly on whether the training data contains information relevant to the research question (e.g., McCoy et al., 2023; Santurkar et al., 2023). We therefore recommend researchers critically assess to what degree the training data—in general—can inform a research question. For example, research on customer satisfaction or service failures may have a relevant representation in the training data (e.g., through data from review platforms) while research on a very specific target audience (e.g., users of only a specific brand) may not be appropriately captured. Identifying the population to which the results generalize is a fundamental challenge in this regard. Training data sets, such as *The Pile*, are simply a collection of massive amounts of data without a clearly defined population. Since the population remains undefined, researchers need to make their prompts more concrete to generate target group-specific results (Argyle et al., 2023), assuming that these target groups are adequately represented in the training data. However, this endeavor's effectiveness depends strongly on the prompt structure, as well as the LLM used, and could vary across versions. For example, GPT-4 demonstrates a higher range of capabilities than GPT-3.5 does with respect to standardized tests (OpenAI, 2023).

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK



Du
Describe differences and similarities between the packages.

ChatGPT
The two packages share several similarities:

1. **Branding**: Both packages prominently feature the brand name "HALOMA" and the word "CHOCOLATE" below it.
2. **Color Scheme**: They both employ a brown and beige color scheme, which is thematic for chocolate products.
3. **Layout**: The positioning of the brand name, logo, net weight (1.55 OZ or 43g), and calorie content (210 calories per bar) is consistent between the two designs.
4. **Logo Design**: The logo, which appears to be a stylized cacao pod or chocolate piece, is identical on both packages.
5. **Minimalist Design**: Despite their differences, both designs lean towards a minimalist aesthetic, without any superfluous text or images beyond the necessary branding and nutritional information.

The differences are in the additional design elements and the overall feel they convey:

1. **Background Detail**: The first package has a plain background, while the second includes a realistic, melting chocolate effect at the top and chocolate splashes, creating a more dynamic and tactile appeal.
2. **Texture and Depth**: The second package's design adds depth and texture through the chocolate imagery, which is absent in the more flat and clean design of the first package.
3. **Visual Impact**: The first package's simplicity gives it a more modern and possibly upscale appearance, while the second package's rich imagery might be seen as more indulgent, appealing to the senses directly.

The second design is more complex and could be intended to attract attention on shelves and evoke the richness of the chocolate, while the first design's simplicity could be aimed at a different market segment that prefers clean and straightforward packaging.

FIGURE 5 Using DALL-E 3 to describe differences and similarities between visual stimuli.

3.2.2 | Customize the LLM and optimize the prompt

Using LLMs to simply interrogate the training data without further customization will probably not yield meaningful results and merely produce a generic response with little variation (Park et al., 2023). Approaches to customization include fine- and prompt-tuning to improve the LLM's performance in a designated task (Demszky et al., 2023). Furthermore, researchers should use probing to assess the output's robustness against, for example, linguistic features (Manning et al., 2020). This process requires first identifying features pertinent to the concept of interest (e.g., negation, use of first person, or synonyms) and varying the input based on these characteristics to

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK

TABLE 1 Guiding questions to consider when using silicon samples in research projects.

Step in the research process	Guiding questions
Alignment of the training data and the research question	*Can the training data inform the research question in a valid way?* Do the training data include information relevant to the research question? Which aspects of the training data could affect the results (e.g., are certain concepts misrepresented)? To which population should the results be generalizable? Is this population sufficiently represented in the training data?
Parametrization, fine-tuning, prompt-tuning, and probing	*How should the silicon sample be generated?* Which settings (e.g., Softmax temperature, top-k) should be used? How should an LLM be additionally fine-tuned for a specific task? How should prompt-tuning be used to draw a silicon sample that is representative of the population of interest? How large should the silicon sample be? Should multiple silicon samples from multiple settings and prompts be combined?
Human benchmarking	*How should the silicon sample be benchmarked with a human sample?* What makes a good human benchmark for the research project? How to deal with deviating results between a human sample and a silicon sample? What are the key quality criteria when comparing human and silicon samples (e.g., equivalent results between the silicon and the human sample)?
Data quality and data analysis	*How does silicon sampling affect established analytical workflows?* What data quality checks should be applied to silicon samples? What are the relevant indicators of data quality? What data-analytical procedures (e.g., significance testing) are appropriate for silicon samples?
Transparency, metadata, and licensing	*What are reasonable reporting standards for silicon samples?* To what granularity do researchers need to report how, when, and where a silicon sample was collected? Which information needs to be provided for other researchers to be able to reproduce the results? Which licenses (distribution and reuse agreements) apply to silicon samples, if any?

examine their influence on the output. For example, researchers can generate a set of sentence pairs as input, differentiated solely by the inclusion or exclusion of negation in verbs (e.g., "satisfied" vs. "not satisfied"). Subsequently, they can analyze the difference in the outputs generated by the model in response to sentences with and without negation. This comparison helps in assessing whether the model accounts for negation in its predictions, and if it does, identifying which elements in the vector show the strongest association with negation (Demszky et al., 2023). Our assessment of order effects regarding different packaging designs (Appendix Figure A2) is another example of a probing task. We recommend that researchers make use of these customization approaches such as iteratively adjusting the prompts if necessary. In doing so, researchers should check the output and, for example, confirm that the output follows the intended format (e.g., selecting one out of multiple options or writing a short text). Under no circumstances, should researchers adjust prompts to generate a specific output content. Ophthalmology researchers Taloni et al. (2023) recently used GPT to create a fake data set that is practically indistinguishable from regular data, but produces false medical evidence. Commenting on their results, one of the authors noted in an interview, "The possibilities are endless, and increasing the quality of the prompts may lead to even more detailed and realistic datasets compared to the one we fabricated" (Fiore, 2023).

When it comes to customizing the model, researchers need to be aware that the cure could be worse than the disease in that their intervention potentially introduces additional biases that extend those that traditional confounding produces. For example, prompt design could be a reflection of researchers' prior beliefs and expectations, thereby unconsciously inducing a confirmation bias. Such practice could easily turn a confirmatory study into an exploratory fishing expedition—a practice that has been criticized as *p*-hacking (Guo & Ma, 2022; Sarstedt & Adler, 2023; Simonsohn et al., 2014) and which is commonly viewed as a major contributor to low replication rates in various fields (Ioannidis, 2005; Miller & Ulrich, 2022). Researchers should also consider enforcing a certain degree of result variability to generate a range of plausible results, rather than striving for a precise estimate that primarily reflects the

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK Copyright Clearance Center

training data's idiosyncrasies and the researcher's degrees of freedom. From this perspective, a certain degree of variability is required to ensure generalizable results. Consequently, researchers should be aware of the potential sensitivity of results to customization approaches. For example, tailoring an LLM to a specific research question might also mean that its outputs are not generalizable, which, in turn, implies that fine-tuned and prompt-tuned models may only be of limited use for hypothesis generation and exploratory research.

3.2.3 | Use a human benchmark sample

In light of the challenges identified above, we recommend that researchers always benchmark their silicon samples with human samples to avoid the risk of tapping into an area where LLMs and humans react differently. Such benchmarking does, of course, offset some of silicon sampling's cost and time advantages. To reap the benefits of silicon sampling, researchers should therefore employ a parsimonious research design for a human benchmarking study, or use secondary data.[5] When conducting such benchmarking, researchers should compare results from different silicon samples (e.g., arising from the use of different LLMs) to identify which one matches the human sample most closely. In doing so, however, care needs to be taken not to overfit the silicon sample to a specific human benchmark in a way that the silicon sample may perfectly mirror one human sample's results but will not generalize to other human samples.

3.2.4 | Justify and adapt analytical procedures

Researchers also need to be aware that silicon samples pose challenges to standard data analytical procedures. For example, researchers could easily produce very large silicon samples that minimize standard errors, thereby rendering standard inference testing of limited value. Similar to big data applications, we recommend researchers abandon significance testing for very large silicon samples and focus on the effect sizes' interpretation (Anderson, 2022) or switch to Bayesian data analysis (Wagenmakers, 2007). However, given the nature of the training data and the sensitivity of the results to the prompt structure, the corresponding estimates are probably associated with a substantial amount of uncertainty, which is quantified doubt about the value of the measurand (i.e., the quantity whose value is sought; JCGM, 2012). High uncertainty implies that a measurement is consistent with a wide range of plausible values for the measurand, which may lead to a wide range of obtained values across different measurements and studies. Quantifying and managing this uncertainty are major challenges—as recent research has highlighted in related contexts (Rigdon & Sarstedt, 2022; Rigdon et al., 2020; Rigdon et al., 2023).

3.2.5 | Optimize reproducibility and transparency

Given the potential sensitivity of the results to various researcher's degrees of freedom, we highly recommend that researchers record as much information as possible (e.g., different variants of prompts; which LLM version was used) and adhere to transparent reporting practices, which allow other researchers to replicate the methodology and reproduce the results. While full reproducibility will hardly be possible due to output variability, enough information should be offered for other researchers to assess, for example, the quality of prompts given the purpose that the silicon sample seeks to fulfill.

3.2.6 | Recommendations

Table 2 summarizes our key recommendations in light of the state-of-research on silicon sampling. Researchers should view the recommendations as a checklist to guide their future research projects employing LLMs.

3.3 | Ethical issues in silicon sampling

Researchers in the fields of psychology and marketing as a whole should also be aware of ethical issues generally raised by LLMs and, specifically, by the use of silicon samples. Ethical issues concern moral judgments such as saying that an action is "right" and "wrong," or "good" and "bad." Ethics as a field of philosophy has proposed various perspectives from which to evaluate ethical judgments, of which deontological and utilitarian ethics are the most common ones. Within applied ethics, the ethics of technology and particularly the ethics of AI and GPT as part of a socio-technical ecosystem have recently emerged as a new area of philosophical inquiry (Stahl et al., 2024; Stahl, 2022). Regarding GPT, general ethical questions include privacy concerns, copyright issues, misinformation and disinformation, bias in training data, job replacement, malicious intent, reinforcement of stereotypes, which impose responsibilities on individual users, and even regulation (https://aicontentfy.com/en/blog/ethical-considerations-of-chatgpt-and-ai). Most of these issues also apply to the use of silicon samples. For example, users should ensure that the generated content is not used to mislead or deceive others; they must respect intellectual property rights; and they should be mindful of privacy concerns. In addition, they are accountable for any errors and misguided implications they communicate. Transparency about the use of LLM-generated content is also key, as it helps maintain trust and integrity in communications (Memarian & Doleck, 2023).

Whether silicon samples are right or wrong, or good or bad in general, and suitable for a particular research question may be

---

[5] Given that many psychological and behavioral effects might not replicate (e.g., Open Science Collaboration, 2015; Röseler et al., 2022), benchmarking silicon samples against human samples data might pose a shifting target. Failures to replicate might not necessarily indicate a silicon sample shortcoming, but can reflect a variation in the original effect's size.

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK

TABLE 2 Key recommendations.

Aspect	Recommendation
Use cases	• Use LLMs and other forms of generative AI (e.g., DALL-E) primarily for tasks whose results can be independently evaluated by researchers. Examples: • Generating items for scale development and index construction • Pretesting survey items • Crafting and pretesting visual stimuli or vignettes • Use silicon samples for research questions, which are likely to have relevant coverage in the training data. Avoid research on very specific target audiences and concepts that are not sufficiently represented in the training data.
Customization, prompting, and probing	• Customize the model by supplying more specific training data (fine-tuning) or customizing the prompts (prompt-tuning). • Apply probing to safeguard the output's robustness against linguistic features or order effects.
Benchmarking	• Benchmark a silicon sample with a relevant human sample. • Evaluate whether the silicon sample matches the human sample in measures for central tendency and variation. • Disregard silicon samples that do not show variation (but provide a single "correct answer" (Park et al., 2023).
Transparency and reproducibility	• Always save a record of model settings, input parameters, and outputs. • Preferably document all relevant information (e.g., model settings, input parameters, and outputs) in repositories such as the OSF to increase transparency.

assessed using the deontological perspective or the utilitarian ethical perspective. From a deontological perspective, silicon samples may be opposed categorically. Some might argue that data from silicon samples are a very different data collection type, which violates research values and a researcher's responsibility to collect original data from human respondents. In comparison, the utilitarian perspective seems to be a more pragmatic perspective considering the costs and benefits of using LLMs.

One major cost discussed in this article concerns the lack of accuracy (or validity) that may turn silicon samples on a grand scale into "silly samples" that simply parrot human texts or worse be misused to forge seemingly "novel" and "interesting," but ultimately misleading findings. In this respect, an established area of research that is well-represented in the training data (e.g., product and service evaluations) may result in more accurate results based on a silicon sample than a less established research area (e.g., consumer response to a crisis such as COVID-19).

Another cost concerns potential biases in existing samples; it is important for researchers to recognize and mitigate biases that may be present in the training data of these models to prevent perpetuation of stereotypes or unfair representations. Researchers have even referred to a new "AI colonialism" due to, for example, the training data painting an incomplete, and potentially biased, picture of non-Western cultures (Hao, 2022), as the data normally originates from Western institutions. Atari et al. (2023) provide support for this notion. The authors contrasted culture-specific beliefs in the World Value Survey with GPT outputs and found that GPT's performance to mimic human responses declines considerably for non-WEIRD (Western, Educated, Industrialized, Rich, and Democratic) versus WEIRD countries. Finally, there is the cost of not sufficiently considering new consumer and marketing-relevant developments and new information. In this case, research topics that are less variant given short-term developments such as customer engagement (e.g., Hollebeek et al., 2024) may be studied more accurately than, for example, research relating to very recent phenomena such as the COVID-19 pandemic (Imschloss & Schwemmle, 2023).

These potential costs need to be compared with some major benefits of silicon samples. Benefits include immense cost and time savings, thereby overcoming inequity issues among researchers. As the number of required empirical studies in consumer and marketing journals has been increasing, critiques have been voiced that some researchers who have fewer funds are being "priced out" of the game. Silicon samples offer a solution to this equity issue.

4 | THE WAY FORWARD

LLMs' ability to deterministically produce correct answers to logic puzzles (e.g., cognitive reflection tasks) and standardized tests is certainly impressive by computer science standards but proves problematic when used to mimic human response behavior with all its imperfections and facets. If LLMs such as GPT act like rational agents, their results cannot readily be used to explain or predict consumers' bounded rational decision-making. Besides the differences in risk preferences and system 1 processing, GPT, specifically, can hardly mimic interpersonal differences, which is crucial for psychological or marketing research studies (e.g., Abdurahman et al., 2023; Park et al., 2023; Santurkar et al., 2023). However, given that the field is evolving rapidly (i.e., industrial players, such as OpenAI and Google, release improved LLMs in quick succession), it is reasonable to assume that future LLM implementations mimic human behavior more closely than current implementations do. For example,

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK

multimodal LLMs interact with external information sources, tools, sensory data, and images, thereby increasing their data richness substantially. Similarly, structured and even semistructured task performances improved substantially after the latest GPT releases, and competitors such as Google's Gemini began outperforming GPTs (Gemini Team Google et al., 2023). We, therefore, advise researchers who consider using silicon samples to take these developments into account because new and improved versions may address some of the pitfalls mentioned in this paper or at least reduce their negative impact.

Great tasks lie ahead: Researchers should not only evaluate LLMs across a broader set of conditions and tasks but also explain LLMs' responses vis-à-vis human decision-making and behavior (see also Binz & Schulz, 2023; Dillion et al., 2023) to identify pathways for using them constructively in scholarly research (Demszky et al., 2023; Susarla et al., 2023).

Commenting on GPT models' suboptimal performance when replicating fundamental behavioral effects, Park et al. (2023, p. 24) note that "such behavioral differences were arguably foreseeable, given that LLMs and humans constitute fundamentally different cognitive systems: with different architectures and potentially substantial differences in the mysterious ways by which each of them has evolved, learned, or been trained to mechanistically process information." Decades of consumer behavior research sought to shed light on these "mysterious ways," but corresponding research in the fields of machine behavior (Rahwan et al., 2019) and machine psychology (Hagendorff et al., 2023) is still nascent. Researchers could interrogate an LLM to identify the reasons for the outputs, but doing so creates a circular problem; Demszky et al. (2023, p. 11) note that: "examining their black box with yet another black box explanation is hardly a reliable approach." Consequently, this approach neither helps identify the LLM's inner working principles nor does it help explain the results of silicon samples.

Studies that compare LLMs' outputs with human response behavior, such as those by Binz and Schulz (2023), Hagendorff et al. (2023), and Park et al. (2023), are an important first step toward identifying human to LLM (dis)similarities, but they also raise a more fundamental question: Do such assessments primarily facilitate understanding mechanisms in LLMs, or do they help researchers understand when and how LLMs mirror human behavior? The two are different endeavors, particularly since the understanding of what constitutes the "correct" response is likely to differ among those who develop LLMs and those who research consumer behavior. An LLM might be "good" for a computer scientist if it tends to produce text that is grammatically sound, coherent, natural-sounding, and, given the input, gives an objective and factually correct answer (Demszky et al., 2023). An LLM-generated answer is correct for a consumer and marketing researcher if it maps actual consumer behavior, considering all its subjectivity. This mapping should not only be equated with predicting behavior but should extend to reflecting on the human reasons that lead to the observed behavior. Infusing the human element into LLMs is a very challenging task and we expect that, in the long run, the market will differentiate between models that excel at providing factually correct outputs and those that produce silicon samples predicting, perhaps even explaining human behavior. The use of silicon samples raises a number of additional questions, ranging from the dimensions and acceptable levels of data quality to the philosophical accounts of the nature of knowledge generation in the age of LLMs.

Researchers should reflect on the research areas that are appropriate for silicon sampling and accumulate corresponding empirical evidence. This endeavor is to identify areas of consumer and marketing research in which silicon samples generally align well with human samples and/or with actual behavior (vs. those in which they do not)—ultimately aiming to substitute or annihilate human benchmarks for certain types of research. As Hamilton et al. (2023) suggest, because LLMs are language-based, they may also be useful for discovering new insights in qualitative data sets rather than just mimicking the themes identified by humans. Similarly, in theory-driven research, silicon samples may be used first to mimic human behavior in line with theoretical predictions. Whether LLMs can be used subsequently to link these findings to theory, or as input for the generation of new theories is yet to be shown—initial research in humanities and archeology, however, gives rise to concern in this regard (Lozić & Štular, 2023). In any case, studies that explore LLMs' capabilities should extend beyond GPT and also consider other models such as Aria, Bard, and Claude 2, which researchers have started to subject to comparative analyses (e.g., Agarwal et al., 2023; Lozić & Štular, 2023; Wu et al., 2023).

Furthermore, silicon sampling is fundamentally different from the data collection and analysis procedures that are currently employed in consumer and marketing research. Researchers therefore need to develop guidelines and methods for silicon sampling, for ethical as well as pragmatic reasons. Most notably, it remains unclear how to define data quality in a silicon sample when it comes to mimicking human behavior with all its imperfections—simply using much data is not a quality signal per se (Hair & Sarstedt, 2021). Should quality be solely defined by the ability of silicon samples to mimic human data patterns? Should researchers define specific "attention" check questions for LLMs that can help identify fraudulent single silicon observations and estimate a sample's overall quality? How does silicon sampling align with the principles of sampling theory? What are outliers in the context of silicon samples and how can researchers identify them? How do model parameters (e.g., Softmax temperature) and fine-tuning impact the silicon sample's quality? These are just some of the emerging issues that will require researchers to rethink some of the basic principles of sampling research.

Relatedly, efforts should be made to develop means to identify silicon samples. Real-world data, for example, often include outliers that may not be present in silicon samples where researchers define variable ranges in their prompts (Taloni et al., 2023). Commenting on the misplaced recommendations provided by their fake data set (Taloni et al., 2023), one of the authors noted in an interview that "we will witness an ongoing tug-of-war between fraudulent attempts to use AI and AI detection systems" (Fiore, 2023). Publishers, journal editors, and reviewers need to be aware of this tug-of-war, which

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK

other fields are already witnessing in the context of paper mills (Candal-Pedreira et al., 2022; Day, 2022; Pérez-Neri et al., 2022).

Consumer researchers, psychologists, computer scientists, and experts from other fields will have to team up to develop LLMs that are optimized for human-like reactions and interactions with users—going beyond fine-tuning and prompt-tuning of current LLMs. As a notable step in this direction, Replika promotes "The AI companion who cares" (https://replika.com/), which is optimized for building an emotional relationship with the user. This purpose is closer to consumer research's strive to understand and predict human behavior than that of LLMs such as GPT, which are optimized to outperform humans in various knowledge domains. Making LLMs more human will require close collaboration between researchers from various fields of scientific inquiry, for example, by translating human behavior into computational functions—an approach that is already persistent, for example, in the choice modeling literature (Gonzalez, 2023).

Beyond these empirical questions, silicon samples, and LLMs more generally, raise new philosophy-of-science questions. For example, researchers have bemoaned that GPT may become "progressively a self-licking lollipop" (Taleb, 2023) as an increasing share of future training data will be comprised of earlier GPT outputs, thereby triggering self-reinforcement. Now imagine a future research world in which the sample of choice, or the sole samples being used, are silicon samples because they have evolved to be low-cost, efficient, and able to mimic human responses to a satisfactory degree. Will this lead to a stagnation in new insight because no "new data" about actual human behavior will be created? Will the findings from such samples converge into homogeneous solutions that lack the variability that characterizes us as humans? Also, can consumer researchers identify and study change over time in consumer behavior in such a world without studying *real* people?

5 | CONCLUSION

Silicon samples hold considerable promise for consumer research as a means to provide human-like data quickly and on a large scale, but can LLMs like GPT serve as "guinea pigbots," as Hutson (2023, p. 121) vividly surmise? By today's standards, LLMs could be useful for settings where researchers collect external feedback to inform further steps in the research process. In such situations, LLM results may give rise to concern and induce researchers to reconsider specific aspects of their project. But as with many disruptive developments, the potential for misuse is real. Leaving fraudulent behavior aside, if adopted without sufficient reflection or for tasks for which LLMs have not been designed, LLMs will likely misinform consumer and marketing researchers—with potentially fatal consequences for the field, which is already under close scrutiny in light of the growing concerns about the replicability of research findings and their lack of relevance for managerial decision-making (Adler et al., 2023; Krefeld-Schwalb & Scheibehenne, 2023). Given that silicon samples offer considerable potential that may attract their premature use in research, scientists should keep in mind, that "with great power comes great responsibility."

ACKNOWLEDGMENTS

Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the second author upon request.

ORCID

Marko Sarstedt http://orcid.org/0000-0002-5424-4268

REFERENCES

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., & Dehghani, M. (2023). *Perils and opportunities in using large language models in psychological research*. OSF Preprints https://doi.org/10.31219/osf.io/tg79n

Adler, S. J., Röseler, L., & Schöniger, M. K. (2023). A toolbox to evaluate the trustworthiness of published findings. *Journal of Business Research*, *167*, 114189. https://doi.org/10.1016/j.jbusres.2023.114189

Agarwal, M., Goswami, A., & Sharma, P. (2023). Evaluating ChatGPT-3.5 and Claude-2 in answering and explaining conceptual medical physiology multiple-choice questions. *Cureus*, *15*(9), e46222. https://doi.org/10.7759/cureus.46222

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). *Using large language models to simulate multiple humans and replicate human subject studies*. arXiv. http://arxiv.org/pdf/2208.10264v5

Anderson, B. S. (2022). What executives get wrong about statistics: Moving from statistical significance to effect sizes and practical impact. *Business Horizons*, *65*(3), 379–388. https://doi.org/10.1016/j.bushor.2021.05.001

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

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023). *Which humans?* arXiv. https://doi.org/10.31234/osf.io/5b26t

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Brand, J., Israeli, A., & Ngwe, D. (2023). *Using GPT for market research*. (Harvard Business School Marketing Unit Working Paper No. 23-062). SSRN. https://doi.org/10.2139/ssrn.4395751

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). *Language models are few-shot learners*. arXiv. http://arxiv.org/abs/2005.14165

Candal-Pedreira, C., Ross, J. S., Ruano-Ravina, A., Egilman, D. S., Fernández, E., & Pérez-Ríos, M. (2022). Retracted papers originating from paper mills: Cross sectional study. *BMJ*, *379*, e071517. https://doi.org/10.1136/bmj-2022-071517

Caron, G., & Srivastava, S. (2022). *Identifying and manipulating the personality traits of language models*. arXiv. http://arxiv.org/abs/2212.10276

Carstensen, L. L. (2006). The influence of a sense of time on human development. *Science*, *312*(5782), 1913–1915. https://doi.org/10.1126/science.1127488

Chang, C.-C., Reitter, D., Aksitov, R., & Sung, Y.-H. (2023). *KL-divergence guided temperature sampling*. arXiv. http://arxiv.org/abs/2306.01286

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK

Chen, Y., Andiappan, M., Jenkin, T., & Ovchinnikov, A. (2023). *A manager and an AI walk into a bar: Does ChatGPT make biased decisions like we do?* SSRN. https://doi.org/10.2139/ssrn.4380365

Day, A. (2022). Exploratory analysis of text duplication in peer-review reveals peer-review fraud and paper mills. *Scientometrics*, *127*(10), 5965–5987. https://doi.org/10.1007/s11192-022-04504-5

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*, 688–701. https://doi.org/10.1038/s44159-023-00241-5

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Earley, P. C., & Ang, S. (2003). *Cultural intelligence: Individual interactions across cultures*. Stanford University Press.

Fiore, K. (2023, November 9). *ChatGPT created a fake dataset with skewed results*. MedpageToday. https://www.medpagetoday.com/special-reports/features/107247

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). *The Pile: An 800GB dataset of diverse text for language modeling*. arXiv. https://arXiv.org/pdf/2101.00027.pdf

Gemini Team Google, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Petrov, S., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., & Vinyals, O. (2023). *Gemini: A family of highly capable multimodal models*. arXiv. http://arxiv.org/abs/2312.11805

Gonzalez, C. (2023). Building human-like artificial agents: A general cognitive algorithm for emulating human decision-making in dynamic environments. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/17456916231196766

Guo, D., & Ma, Y. (2022). The "p-hacking-is-terrific" ocean - A cartoon for teaching statistics. *Teaching Statistics*, *44*(2), 68–72. https://doi.org/10.1111/test.12305

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, *3*, 833–838. https://doi.org/10.1038/s43588-023-00527-x

Hair, J. F., & Sarstedt, M. (2021). Data, measurement, and causal inferences in machine learning: Opportunities and challenges for marketing. *Journal of Marketing Theory and Practice*, *29*(1), 65–77. https://doi.org/10.1080/10696679.2020.1860683

Hamilton, L., Elliott, D., Quick, A., Smith, S., & Choplin, V. (2023). Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. *International Journal of Qualitative Methods*, *22*, 1–13. https://doi.org/10.1177/16094069231201504

Hao, K. (2022, April 19). *Artificial intelligence is creating a new colonial world order*. MIT Technology Review. https://www.technologyreview.com/2022/04/19/1049592/artificial-intelligence-colonialism/

Hollebeek, L. D., Menidjel, C., Sarstedt, M., Jansson, J., & Urbonavicius, S. (2024). Engaging consumers through artificially intelligent technologies: Systematic review, conceptual model, and further research. *Psychology & Marketing*. Advance online publication. https://doi.org/10.1002/mar.21957

Hutson, M. (2023). Guinea pigbots. *Science*, *381*(6654), 121–123. https://doi.org/10.1126/science.adj6791

Imschloss, M., & Schwemmle, M. (2023). Value creation in post-pandemic retailing: A conceptual framework and implications. *Journal of Business Economics*. Advance online publication. https://doi.org/10.1007/s11573-023-01189-x

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

JCGM. (2012). *International vocabulary of metrology—Basic and general concepts and associated terms (VIM)* (3 ed.https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf

Jeon, J., Lee, S., & Choi, S. (2023). A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models. *Interactive Learning Environments*. Advance online publication. https://doi.org/10.1080/10494820.2023.2204343

Jiang, H., Zhang, X., Cao, X., & Kabbara, J. (2023). *PersonaLLM: Investigating the ability of large language models to express Big Five personality traits*. arXiv. http://arxiv.org/abs/2305.02547

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292.

Kirshner, S. N. (2024). GPT and CLT: The impact of ChatGPT's level of abstraction on consumer recommendations. *Journal of Retailing and Consumer Services*, *76*, 103580. https://doi.org/10.1016/j.jretconser.2023.103580

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Sowden, W. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Krefeld-Schwalb, A., & Scheibehenne, B. (2023). Tighter nets for smaller fishes? Mapping the development of statistical practices in consumer research between 2008 and 2020. *Marketing Letters*, *34*(3), 351–365. https://doi.org/10.1007/s11002-022-09662-3

Lampinen, A. K., Dasgupta, I., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2023). *Language models show human-like content effects on reasoning tasks*. arXiv. http://arxiv.org/abs/2207.07051

Lietz, P. (2010). Research into questionnaire design. *International Journal of Market Research*, *52*(2), 249–272.

Lozić, E., & Štular, B. (2023). *ChatGPT v Bard v Bing v Claude 2 v Aria v human-expert. How good are AI chatbots at scientific writing?* arXiv. https://doi.org/10.3390/fi15100336

Luo, B., Lau, R. Y. K., Li, C., & Si, Y. W. (2022). A critical review of state-of-the-art chatbot designs and applications. *WIREs Data Mining and Knowledge Discovery*, *12*(1):e1434. https://doi.org/10.1002/widm.1434

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). *Embers of autoregression: Understanding large language models through the problem they are trained to solve*. arXiv. https://arXiv.org/pdf/2309.13638.pdf

Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, *5*, 100152. https://doi.org/10.1016/j.caeai.2023.100152

Miller, J., & Ulrich, R. (2022). Optimizing research output: How can psychological research methods be improved? *Annual Review of Psychology*, *73*, 691–718. https://doi.org/10.1146/annurev-psych-020821-094927

Ooi, K.-B., Tan, G. W.-H., Al-Emran, M., Al-Sharafi, M. A., Capatina, A., Chakraborty, A., Dwivedi, Y. K., Huang, T.-L., Kar, A. K., Lee, V.-H., Loh, X.-M., Micu, A., Mikalef, P., Mogaji, E., Pandey, N., Raman, R., Rana, N. P., Sarker, P., Sharma, A., … Wong, L. W. (2023). The

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK

potential of generative artificial intelligence across disciplines: Perspectives and future directions. *Journal of Computer Information Systems*. Advance online publication. https://doi.org/10.1080/08874417.2023.2261010

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

OpenAI. (2023). *GPT-4 technical report*. arXiv. https://arxiv.org/pdf/2303.08774

Park, P. S., Schoenegger, P., & Zhu, C. (2023). *Diminished diversity-of-thought in a standard large language model*. arXiv. https://arxiv.org/pdf/2302.07267.pdf

Paul, J., Ueno, A., & Dennis, C. (2023). ChatGPT and consumers: Benefits, pitfalls and future research agenda. *International Journal of Consumer Studies*, *47*(4), 1213–1225. https://doi.org/10.1111/ijcs.12928

Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing*, *40*(2), 269–275. https://doi.org/10.1016/j.ijresmar.2023.03.001

Pérez-Neri, I., Pineda, C., & Sandoval, H. (2022). Threats to scholarly research integrity arising from paper mills: A rapid scoping review. *Clinical Rheumatology*, *41*(7), 2241–2248. https://doi.org/10.1007/s10067-022-06198-9

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Rigdon, E., Sarstedt, M., & Moisescu, O. I. (2023). Quantifying model selection uncertainty via bootstrapping and Akaike weights. *International Journal of Consumer Studies*, *47*(4), 1596–1608. https://doi.org/10.1111/ijcs.12906

Rigdon, E. E., Sarstedt, M., Becker, J. M., (2020). Quantify uncertainty in behavioral research. *Nature Human Behaviour*, *4*(4), 329–331. https://doi.org/10.1038/s41562-019-0806-0

Rigdon, E. E., & Sarstedt, M. (2022). Accounting for uncertainty in the measurement of unobservable marketing phenomena. In H. Baumgartner, & B. Weijters (Eds.), *Review of marketing research* (19th ed., pp. 53–73). Emerald.

Röseler, L., Doetsch, C., Kaiser, L., Klett, N., Krapp, J., Seida, J., Schütz, C., Cummins, A., Dienlin, J., Elsherif, T., Förster, M., Genschow, N., Gnambs, O., Hartmann, T., Holgado, H., Hussey, D., Korbmacher, M., Kulke, L., Liu, Y., Lou, N., & Vaughn, L. (2022). *ReD: Replication Database*. https://doi.org/10.17605/OSF.IO/9r62x

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). *Whose opinions do language models reflect?* arXiv. https://arxiv.org/pdf/2303.17548

Sarstedt, M., & Adler, S. J. (2023). An advanced method to streamline p-hacking. *Journal of Business Research*, *163*, 113942. https://doi.org/10.1016/j.jbusres.2023.113942

Schlägel, C., & Sarstedt, M. (2016). Assessing the measurement invariance of the four-dimensional cultural intelligence scale across countries: A composite model approach. *European Management Journal*, *34*(6), 633–649. https://doi.org/10.1016/j.emj.2016.06.002

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Stahl, B. C. (2022). From computer ethics and the ethics of AI towards an ethics of digital ecosystems. *AI and Ethics*, *2*, 65–77. https://doi.org/10.1007/s43681-021-00080-1

Stahl, B. C., & Eke, D., Stahl and Eke. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, *74*, 102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700

Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus Effect of generative AI: Charting the path for responsible conduct of scholarly activities in information systems. *Information Systems Research*, *34*(2), 399–408. https://doi.org/10.1287/isre.2023.ed.v34.n2

Taleb, N. N. (2023). ChatGPT is a statistical representation of things found on the web, which will increasingly include ITS OWN output (directly and second hand). You post something picked up from it & it will use it to reinforce its own knowledge. Progressively a self-licking lollipop. Enjoy #AI [@nntaleb] [Tweet]. Twitter. https://twitter.com/nntaleb/status/1660304315217584128

Taloni, A., Scorcia, V., & Giannaccare, G. (2023). Large language model advanced data analysis abuse to create a fake data set in medical research. *JAMA Ophthalmology*, *141*, 1174. https://doi.org/10.1001/jamaophthalmol.2023.5162

Ton, L. A. N., Smith, R. K., & Sevilla, J. (2023). Symbolically simple: How simple packaging design influences willingness to pay for consumable products. *Journal of Marketing*. Advance online publication. https://doi.org/10.1177/00222429231192049

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

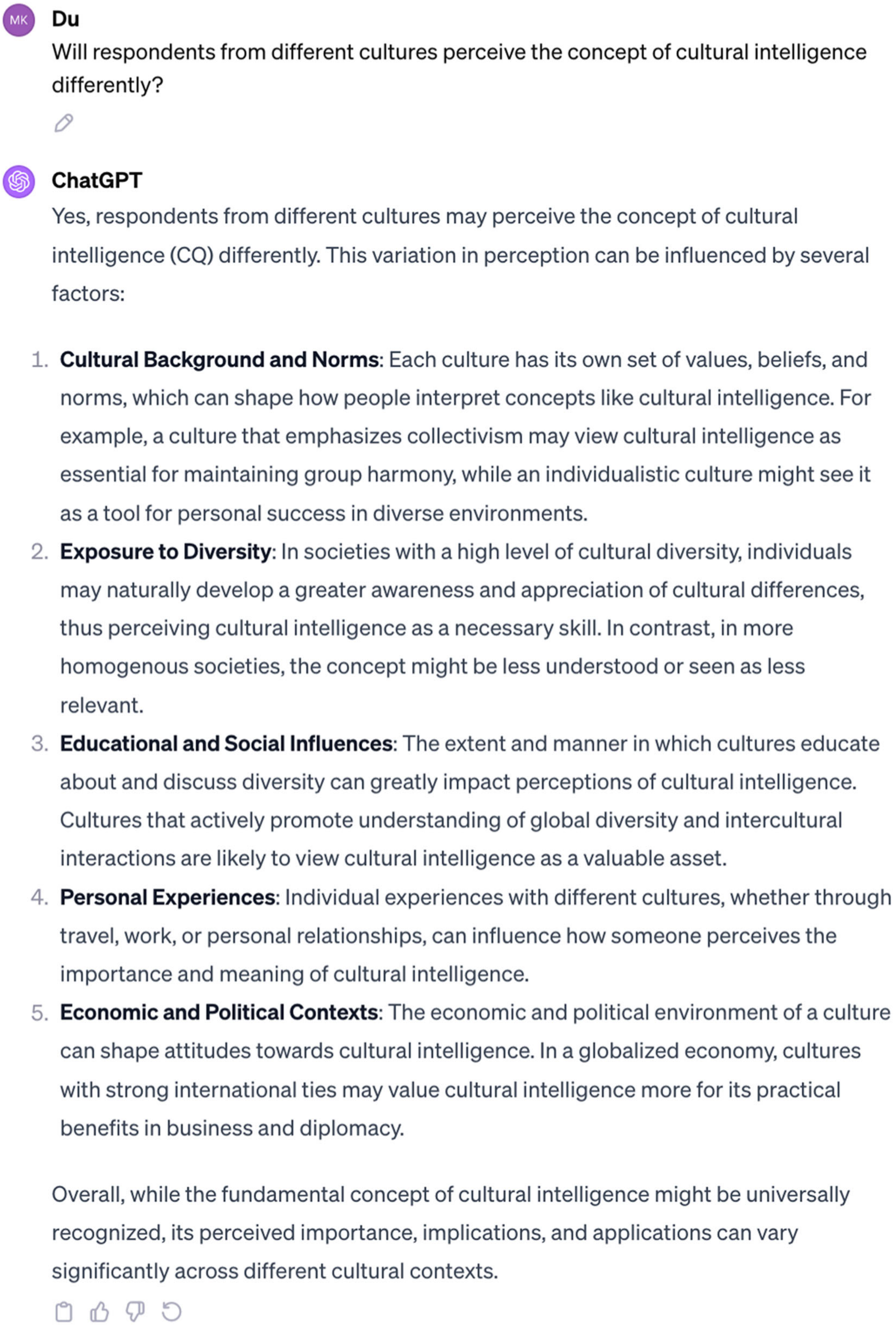
Wu, S., Koo, M., Blum, L., Black, A., Kao, L., Scalzo, F., & Kurtz, I. (2023). *A comparative study of open-source large langu age models, GPT-4 and Claude 2: Multiple-choice test taking in nephrology*. arXiv. https://arXiv.org/abs/2308.04709

How to cite this article: 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*, 1254–1270. https://doi.org/10.1002/mar.21982

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK Copyright Clearance Center

APPENDIX



Du

Will respondents from different cultures perceive the concept of cultural intelligence differently?

ChatGPT

Yes, respondents from different cultures may perceive the concept of cultural intelligence (CQ) differently. This variation in perception can be influenced by several factors:

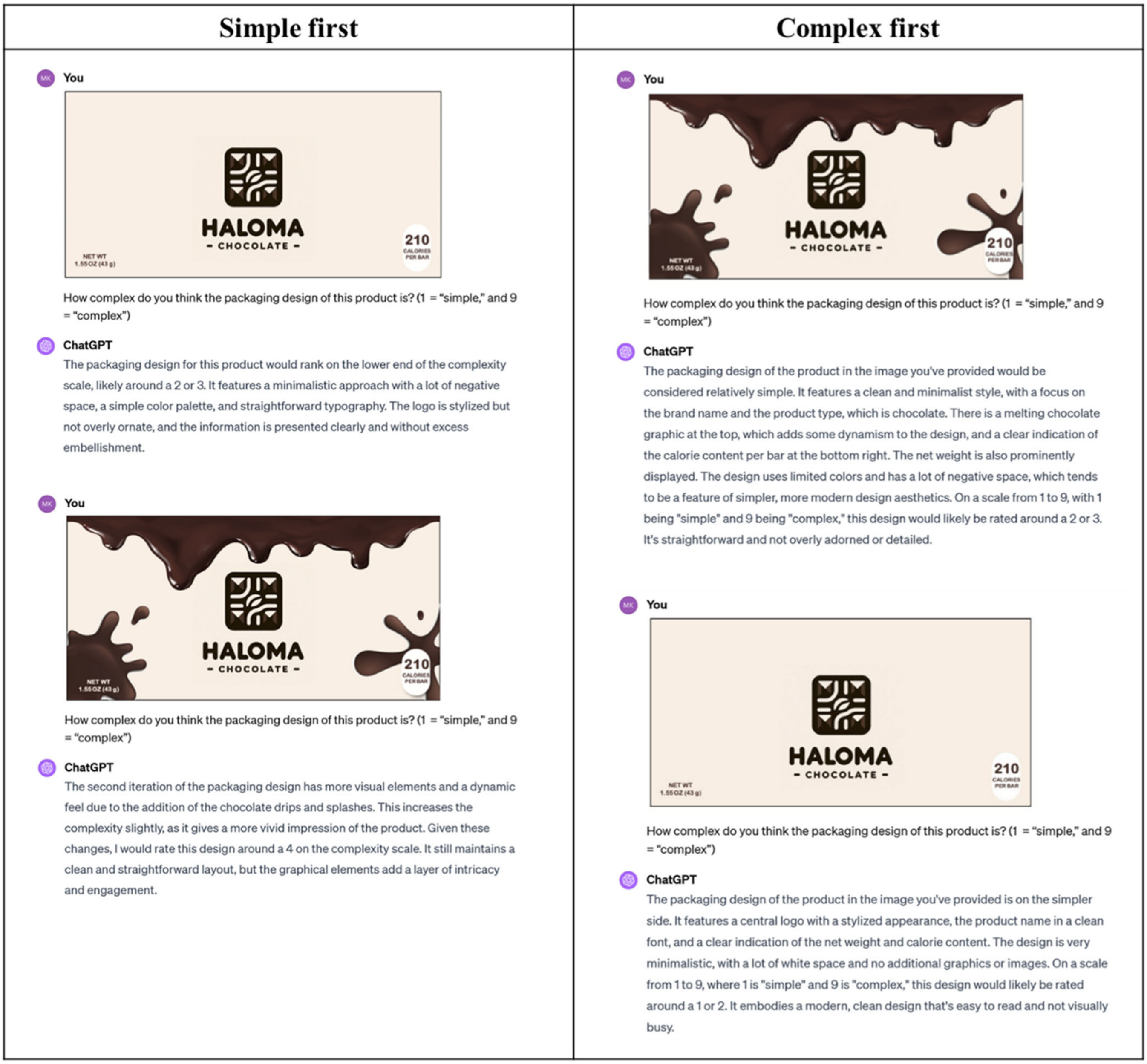
1. **Cultural Background and Norms**: Each culture has its own set of values, beliefs, and norms, which can shape how people interpret concepts like cultural intelligence. For example, a culture that emphasizes collectivism may view cultural intelligence as essential for maintaining group harmony, while an individualistic culture might see it as a tool for personal success in diverse environments.
2. **Exposure to Diversity**: In societies with a high level of cultural diversity, individuals may naturally develop a greater awareness and appreciation of cultural differences, thus perceiving cultural intelligence as a necessary skill. In contrast, in more homogenous societies, the concept might be less understood or seen as less relevant.
3. **Educational and Social Influences**: The extent and manner in which cultures educate about and discuss diversity can greatly impact perceptions of cultural intelligence. Cultures that actively promote understanding of global diversity and intercultural interactions are likely to view cultural intelligence as a valuable asset.
4. **Personal Experiences**: Individual experiences with different cultures, whether through travel, work, or personal relationships, can influence how someone perceives the importance and meaning of cultural intelligence.
5. **Economic and Political Contexts**: The economic and political environment of a culture can shape attitudes towards cultural intelligence. In a globalized economy, cultures with strong international ties may value cultural intelligence more for its practical benefits in business and diplomacy.

Overall, while the fundamental concept of cultural intelligence might be universally recognized, its perceived importance, implications, and applications can vary significantly across different cultural contexts.

FIGURE A1 Measurement invariance testing.

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK



Simple first

You

How complex do you think the packaging design of this product is? (1 = "simple," and 9 = "complex")

ChatGPT

The packaging design for this product would rank on the lower end of the complexity scale, likely around a 2 or 3. It features a minimalistic approach with a lot of negative space, a simple color palette, and straightforward typography. The logo is stylized but not overly ornate, and the information is presented clearly and without excess embellishment.

You

How complex do you think the packaging design of this product is? (1 = "simple," and 9 = "complex")

ChatGPT

The second iteration of the packaging design has more visual elements and a dynamic feel due to the addition of the chocolate drips and splashes. This increases the complexity slightly, as it gives a more vivid impression of the product. Given these changes, I would rate this design around a 4 on the complexity scale. It still maintains a clean and straightforward layout, but the graphical elements add a layer of intricacy and engagement.

Complex first

You

How complex do you think the packaging design of this product is? (1 = "simple," and 9 = "complex")

ChatGPT

The packaging design of the product in the image you've provided would be considered relatively simple. It features a clean and minimalist style, with a focus on the brand name and the product type, which is chocolate. There is a melting chocolate graphic at the top, which adds some dynamism to the design, and a clear indication of the calorie content per bar at the bottom right. The net weight is also prominently displayed. The design uses limited colors and has a lot of negative space, which tends to be a feature of simpler, more modern design aesthetics. On a scale from 1 to 9, with 1 being "simple" and 9 being "complex," this design would likely be rated around a 2 or 3. It's straightforward and not overly adorned or detailed.

You

How complex do you think the packaging design of this product is? (1 = "simple," and 9 = "complex")

ChatGPT

The packaging design of the product in the image you've provided is on the simpler side. It features a central logo with a stylized appearance, the product name in a clean font, and a clear indication of the net weight and calorie content. The design is very minimalistic, with a lot of white space and no additional graphics or images. On a scale from 1 to 9, where 1 is "simple" and 9 is "complex," this design would likely be rated around a 1 or 2. It embodies a modern, clean design that's easy to read and not visually busy.

FIGURE A2 Order effect in the packaging design evaluation.

15206793, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/mar.21982 by Universitaet Bayreuth, Wiley Online Library on [22/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK