RATIONALITY & COMPETITION
CRC TRR 190

Re-fielding Replicability of Silicon Sample-based Marketing Research

Steffen Jahn (Martin Luther University Halle-Wittenberg)
Tobias Klinke (experial Research Lab)
Daniel Guhl (HU Berlin)
Maja Murr (HU Berlin)

Discussion Paper No. 583

August 05, 2026

Re-fielding Replicability of Silicon Sample-based Marketing Research

Steffen Jahn[a,*], Tobias Klinke[b], Daniel Guhl[c], and Maja Murr[c]

August 2026

Large language models are increasingly used to generate silicon samples, synthetic respondents that stand in for human participants in behavioral research. As silicon samples enter the methodological mainstream, the question of their *re-fielding replicability* becomes critical: when researchers regenerate a silicon sample under the same procedure at a later point in time, how stable are the resulting data? Existing validation studies have compared LLM-generated responses to human responses at a single point in time, but the temporal consistency of silicon samples themselves (whether independent reapplications of the same generation procedure yield comparable samples) has not been examined systematically, especially under sparse-conditioning, re-fielded sample-generation settings. We address this gap by generating silicon samples across three weekly waves, permitting a re-fielding consistency assessment. We benchmark this assessment with a human reference sample recruited via Prolific. Moreover, we systematically compare three structurally distinct generation paradigms that try to increase response variability: single-prompt batch generation, iterative generation with cumulative memory, and two-stage seed-and-expand generation. We find that re-fielding consistency is reasonably high in aggregate and comparable to the human benchmark for the strongest paradigm (two-stage seed-and-expand) but markedly lower for single-prompt batch generation. In contrast, alignment with the human reference is generally low, so that silicon samples reproduce their own outputs more faithfully than they reproduce human data. This means that silicon samples can be internally reproducible without being externally valid.

Keywords: Silicon samples; Synthetic data; Replicability; Consistency; Sustainability; Plant-based meat

JEL: C83, C52, C45, M31, Q56, D12

[a] School of Economics and Business, Martin Luther University Halle-Wittenberg, Universitätsplatz 10, 06108 Halle (Saale), Germany.

[*] Corresponding Author: steffen.jahn@wiwi.uni-halle.de

[b] experial Research Lab, Hohenstaufenring 62, 50674 Cologne, Germany

[c] School of Business and Economics, Humboldt University Berlin, Unter den Linden 6, 10099 Berlin, Germany

1. Introduction

Generative AI is disrupting not only traditional business operations (Hartmann, Exner, & Domdey 2025; Kumar et al. 2025) but also the methods used in behavioral research (Peres et al. 2023; Xie et al. 2026). A particularly promising development for both academics and practitioners is the use of synthetic respondents in surveys, conjoint analyses, and experiments (Sarstedt et al. 2026; Wang, Zhang, & Zhang 2026). Large language models (LLMs) can scale data generation by producing silicon samples to simulate real consumer behavior in a resource-efficient way (Toubia et al. 2025; Wang, Zhang, & Zhang 2026; Yoo, Haenlein, & Hewett 2025).

If we consider silicon samples as substitutes for human participants to save time and costs, a critical question arises: how much should we trust the insights derived from such data? There is an emerging literature that calculates the discrepancies between synthetic data and human responses in cross-sectional settings (Cui, Li, & Zhou 2025; Goli & Singh 2024; Xie et al. 2026). A common finding is that silicon samples perform reasonably well in some areas but align rather poorly with well-established consumer behavior effects (Goli & Singh 2024; Sarstedt et al. 2026). Poor performance is often attributed to LLMs' tendency to overestimate effect sizes (Cui, Li, & Zhou 2025) as well as generate data that concentrates around "typical" values, thus failing to capture real-world heterogeneity (Park, Schoenegger, & Zhu 2024; Xie et al. 2026). In turn, marketing researchers have expressed deep concerns about the unpredictability of LLM-generated responses (Joerling 2026).

A particularly important yet neglected issue pertaining to this unpredictability is the

re-fielding replicability of insights from studies involving silicon samples. Specifically, one might wonder whether an LLM produces stable responses when researchers re-generate a silicon sample under the same procedure at a later point in time. In a quick self-test, as a motivating anecdote, asking *Gemini* to provide the average American consumer's maximum willingness to pay for a ready-to-eat plant-based burger revealed $13.50, $5.50, and $6.75, respectively, across three sessions. If feeding the same prompt to an LLM repeatedly yields responses that vary more than those of human samples, this would limit silicon samples' usefulness for research settings that rely on stable repeated measurement. Additionally, low replicability would make any side-by-side comparison between human and silicon samples seem completely arbitrary. Therefore, tracking the stability of silicon samples over time is just as crucial as replicating traditional human behavioral studies (Nosek et al. 2022).

This paper contributes to the evolving conversation about the benefits and risks of using silicon samples in behavioral research. While previous studies have primarily examined how well an LLM-generated dataset reproduces key statistical properties of corresponding human survey data (Cui, Li, & Zhou 2025; Goli & Singh 2024; Sarstedt et al. 2024; Xie et al. 2026), our work extends the existing literature by quantifying *re-fielding consistency*, i.e., whether re-running the same survey as repeated cross-sections yields similar results. Interestingly, one study has taken a complementary route by comparing simulated with actual responses, adjusting for human participants' self-consistency in a repeated-measures design (Park et al. 2024). In another highly controlled setting, Toubia et al. (2025) evaluated the test-retest accuracy of digital twins trained on a rich dataset, reporting 82% accuracy. In contrast, our study prioritizes realism. We examine population-level replicability across

re-generated samples using sparse conditioning (Xie et al. 2026), an approach that more closely mirrors standard practice in traditional marketing research.

We collected data across three waves (total N = 11,648), spanning three LLMs and three diversity-enforcing paradigms, including "single-prompt batch generation" (Argyle et al. 2023), "iterative generation with cumulative memory" (Alismail & Lanquillon 2025) to generate personas within an agentic workflow as well as "two-stage seed-and-expand generation" (Wang et al. 2023) to generate personas sequentially until a diversity criterion is fulfilled. To benchmark the replicability of silicon samples against human samples, we additionally collected three waves on Prolific. The design enables the assessment of re-fielding consistency, between-wave variability, and response patterns specific to a generation paradigm.

To provide context for our analysis, we administered a survey investigating perceptions of and preferences for plant-based meat alternatives (PBMAs). PBMAs represent a growing market that still faces significant challenges to general acceptance (Jahn, Furchheim, & Strässner 2021; Jahn, Guhl, & Erhard 2024; Statista 2025), making it a particularly interesting research topic from both an academic and managerial perspective. Further, sustainable food preferences represent a sensitive topic, creating a challenging environment for synthetic data (Cui, Li, & Zhou 2025). In this survey, we include questions on demographics, consumer characteristics, food product perceptions, and preferences for animal-based and plant-based food to assess the domains in which silicon samples (do not) replicate well.

Our results show that the re-fielding replicability of silicon samples is reasonably high in aggregate, even under sparse conditioning. Consistency across waves is high on

demographics and preferences, but degrades substantially on consumer characteristics, product perception measures, the internal consistency of multi-item scales, and relational properties such as correlations. Results further suggest that *single-prompt batch* generation produces less consistent data, while *two-stage seed-and-expand* generation, in aggregate, is as consistent as the human benchmark. In line with a more skeptical view (Sarstedt et al. 2026; Xie et al. 2026), we find that silicon samples are not generally aligned with the human reference on the substantive content that motivates their use in marketing research.

2. Conceptual Background

2.1. Silicon Samples in Behavioral Research

The idea that LLMs can stand in for human respondents in survey-based research has moved from a speculative possibility to an active methodological program over the last couple of years. The premise, formalized by Argyle et al. (2023) under the label of algorithmic fidelity, is that LLMs trained on large corpora of human-generated text implicitly encode the distribution of human attitudes and behaviors, and that this distribution can be conditioned on demographic backstories to produce so-called silicon samples that approximate the responses of targeted human subpopulations.

The empirical literature evaluating this premise has produced mixed results. On the optimistic side, Brand, Israeli, & Ngwe (2023) demonstrate that GPT-3.5 produces conjoint-style willingness-to-pay estimates that are broadly consistent with documented patterns in consumer demand, including downward-sloping demand curves and brand-loyalty effects.

Li et al. (2024) report agreement rates above 75% between LLM-generated and human-survey-based brand similarity measures, and Mei et al. (2024) show that ChatGPT replicates qualitative findings in classical economic games such as the ultimatum and trust games. In a particularly large-scale demonstration, Hewitt et al. (2024) prompt GPT-4 to simulate responses to 476 treatment effects drawn from 70 pre-registered survey experiments and report a correlation of $r = 0.85$ between simulated and observed treatment effects—including for studies published after the model's training cutoff. Park et al. (2024), working with individually-conditioned generative agents constructed from two-hour qualitative interviews with > 1,000 participants, report that the agents reproduce participants' General Social Survey responses with 85% of the accuracy of the participants' own two-week test-retest consistency. Cui, Li, & Zhou (2025), replicating 156 scenario-based experiments from leading social science journals across three frontier LLMs, find that 73 to 81 percent of main effects are reproduced in direction and statistical significance.

At the same time, a growing body of research documents systematic discrepancies between LLM-generated and human responses that complicate the optimistic reading. In marketing, systematic guides recommend that the technology is particularly suited to upstream tasks such as research design, pretesting and study administration, but not yet a substitute for human data collection in primary research (Blanchard et al. 2025; Peres et al. 2023; Sarstedt et al. 2024; Yoo, Haenlein, & Hewett 2025). This view pertains particularly to silicon samples under sparse conditioning settings (Wang, Zhang, & Zhang 2026; Xie et al. 2026).

The most consistently reported finding is a reduction in response variance: while LLM outputs may track human means reasonably well in the aggregate, their distributions

are typically narrower, with smaller standard deviations and a tendency for responses to concentrate around modal categories (Bisbee et al. 2024; Park, Schoenegger, & Zhu 2024; Xie et al. 2026). Bisbee et al. (2024), in a large-scale replication of the American National Election Survey with GPT-generated synthetic respondents, find that this variance compression leads to inaccurate subgroup means, attenuated regression coefficients, and in several instances, coefficient estimates of opposite sign to those obtained from the human benchmark. Park, Schoenegger, & Zhu (2024) observe a "correct answer" effect characterized by the LLM answering some questions in a highly uniform way because it treats these questions as if there was a correct answer. A byproduct of these tendencies is that LLMs systematically misrepresent the opinions of certain demographic subgroups, including older adults and politically conservative respondents, even when explicit demographic conditioning is applied (Santurkar et al. 2023). Goli & Singh (2024) report related findings for intertemporal preference elicitation, observing that LLM-generated discount-rate estimates deviate systematically from human benchmarks and respond in linguistically biased ways to the framing of the elicitation. Cui, Li, & Zhou (2025), while finding high replication rates for main effects, also report effect sizes that are two to three times larger than those observed in human studies, suggesting that LLMs amplify rather than faithfully reproduce treatment effects.

The mechanisms underlying these discrepancies are increasingly well understood. Recent work in the machine-learning literature has identified mode collapse, the tendency of preference-aligned models to concentrate their output distribution on the most typical response, as an inherent property of how contemporary LLMs are trained, rather than being an idiosyncrasy of any particular model (Xie et al. 2026; Zhang et al. 2025). Gui &

Toubia (2023) identify a parallel issue from a causal-inference perspective: when LLM respondents are kept blind to the experimental design in the manner standard for human participants, treatment variations systematically affect unspecified variables that should remain constant, violating the unconfoundedness assumption on which causal inference depends. Brucks & Toubia (2025), in a full-factorial analysis of prompt-architecture effects, demonstrate that seemingly arbitrary features of how a prompt is constructed, such as the order in which response options are presented, the labels assigned to them, or the framing of justifications, produce systematic methodological artifacts large enough to overturn substantive conclusions. Taken together, these findings suggest that the apparent demographic differentiation observed in "optimistic" studies may, in fact, be rather fragile and highly dependent on the specific architecture of the prompt.

A more fundamental question for any measurement methodology, however, has so far remained outside the scope of the empirical literature: whether silicon samples are themselves stable across independent reapplications of the same generation procedure. To this point, only one study has examined the temporal stability of silicon samples (Toubia et al. 2025). Based on 2,058 responses to more than 500 questions in a two-and-a-half-hour survey (administered across four waves), Toubia et al. (2025) created digital twins to examine how well they reproduced the original data. As part of this study, 17 tasks were administered twice to participants, allowing an assessment of test-retest accuracy for each digital twin, which could then be aggregated into an accuracy index. The complementary question—how stable the synthetic samples themselves would be if regenerated under the same procedure at a later point in time (a property we call re-fielding consistency), and how such stability relates to alignment with a human reference—has, to our knowledge, not

been addressed systematically. This question is logically prior to the comparison against a human reference: if independent reapplications of the same procedure yield substantively different samples, then any single-shot comparison to a human reference captures only one realization of an underlying distribution of possible samples. This aspect becomes especially relevant when silicon samples are generated in ways that broaden the narrow distributions typically found in synthetic data. As we will argue in the following section, the available generation paradigms are not interchangeable, and the ways they differ in their handling of within-sample diversity have substantive implications for the data they produce, potentially affecting temporal stability.

2.2. Sources of Heterogeneity in Silicon-Sample Generation

The findings reviewed above converge on a structural observation that has not received the theoretical attention it deserves. LLMs, in their default operating mode, do not produce diverse samples. The mechanisms documented in the previous section, mode collapse driven by typicality bias in preference data (Zhang et al. 2025), the correct-answer effect (Park, Schoenegger, & Zhu 2024), the reduced response variance documented across multiple replication attempts (Bisbee et al. 2024; Santurkar et al. 2023), and the systematic misrepresentation of subgroups (Santurkar et al. 2023), all describe variations on a single underlying tendency: a contemporary aligned LLM, when prompted to produce many independent respondent answers, will produce fewer distinct answers than a comparable human sample would. This is not a bug to be eliminated through better prompting, but a consequence of how preference optimization and alignment shape the output distribution of frontier models (Zhang et al. 2025). Diversity, in synthetic sampling, has to be added

back in by the researcher.

This recognition has produced a methodological response that is, by now, the de facto standard in applied silicon-sampling work: persona conditioning. Rather than asking a single model to produce many respondents from an unspecified population, researchers prompt the model with explicit demographic, attitudinal, or biographical descriptions intended to anchor each generated response in a distinct individual perspective (Argyle et al. 2023; Sun et al. 2024). The expectation is that conditioning on persona-level variation will propagate into response-level variation, producing the heterogeneity that the unconditioned model cannot produce on its own. The empirical evidence on whether this expectation is met is, however, considerably weaker than the prevalence of the technique suggests. Hu & Collier (2024) demonstrate that the persona effect, the share of human response variance recovered by persona-prompted LLM simulations, is linearly tied to the strength of the correlation between persona variables and human annotations in the underlying data. For constructs where demographic variables are strongly predictive, persona prompting recovers a substantial share of the variance. However, for the subjective, attitudinal, or behavioral constructs typical to consumer research, the benefit of persona prompting is small. Li et al. (2025), in a recent large-scale evaluation, report a related caution: enriching personas with additional content (e.g., backstories, personality descriptions, and other LLM-generated detail) does not reliably move silicon samples closer to real-world outcomes, and in their setting can move them further away. Together, these findings caution against assuming that the relationship between persona richness and silicon-sample validity is monotonic, or that more detailed personas will, as a general rule, yield more realistic responses.

A second design decision concerns the level at which the persona generation process is anchored to empirical reference data. At one extreme, individual-level replication approaches construct each synthetic respondent based on a specific real person's interview transcript or prior-wave survey responses (Park et al. 2024; Toubia et al. 2025). Distribution-matched approaches anchor the joint demographic distribution of the synthetic sample to that of an empirical reference sample (Argyle et al. 2023; Bisbee et al. 2024; Hewitt et al. 2024). Marginals-based approaches specify only the marginal distributions of selected demographic variables, typically against census targets, without anchoring on a specific empirical sample (Sun et al. 2024). At the opposite extreme, target-group approaches specify only the population eligibility criteria (similar to those used to recruit a human sample) and rely on the model to produce its own internal distribution of personas within those criteria.

These anchoring strategies concern the empirical reference against which a sample is constructed; a separate question is the architecture through which within-sample diversity is enforced. Along this second dimension, three generation paradigms can be distinguished (Alismail & Lanquillon 2025; Chan et al. 2024; Wang et al. 2023). The first, and currently dominant, paradigm treats persona generation as a single-LLM task in which a single model produces all synthetic respondents via repeated, independent prompts. The second paradigm, increasingly prevalent in the broader synthetic-data-generation literature (Alismail & Lanquillon 2025), implements persona generation via an agentic workflow in which multiple LLM calls coordinate via a shared memory structure to iteratively produce respondents. The third paradigm, building on seed-and-expand architectures established in the training-data literature (Chan et al. 2024; Wang et al. 2023), decouples the construc-

tion of a diverse anchor set from the expansion to a full panel through a two-stage pipeline. These three paradigms differ in the structural point at which diversity is enforced in the generation process, and in the operational mechanism by which it is sustained across the resulting sample.

3. Method

3.1. Data Generation

This pre-registered study (`https://aspredicted.org/5683tw.pdf`) uses a repeated cross-sectional design, spanning three waves and multiple data sources: ChatGPT, Qwen, Gemini, Prolific, and Amazon MTurk. To counter LLMs' tendency to generate data that is too homogeneous (Park, Schoenegger, & Zhu 2024; Xie et al. 2026), we implement the three diversity-enforcing generation paradigms described earlier, instantiated in a way intended to bracket realistic researcher practice. The first paradigm, *single-prompt batch generation*, reflects the workflow of a non-technical user interacting with a model through its standard chat interface. The second, *iterative generation with cumulative memory*, reflects the workflow of a technically skilled user who orchestrates an automated, agentic pipeline. The third, *two-stage seed-and-expand generation*, reflects the workflow of a researcher delegating panel construction to a dedicated synthetic-research platform that implements an internal diversity-control loop. The first two paradigms are crossed with three LLM families, *ChatGPT*, *Qwen*, and *Gemini*, yielding six synthetic datasets per wave; the seed-and-expand paradigm contributes one additional dataset, generated with *Gemini* via the default reasoning backbone of the *experial* research platform. The resulting seven synthetic datasets

were produced alongside each of the three weekly waves. Table 1 summarizes the seven synthetic conditions and their model families. We provide detailed information on sample generation for each paradigm in Web Appendix A.

TABLE 1. Overview of the seven synthetic conditions

Generation paradigm	**Sample ID**	**Models**	**Target *n***
Single-prompt batch generation	`ChatGPT single-prompt batch`	*Primary:* `GPT 5.4`	550
Single-prompt batch generation	`Qwen single-prompt batch`	*Primary:* `Qwen 3.6-Plus`	550
Single-prompt batch generation	`Gemini single-prompt batch`	*Primary:* `Gemini 3`	550
Iterative generation with cumulative memory	`ChatGPT iterative`	*Primary:* `gpt-5.4-mini` *Summary:* `gpt-5.4-nano`	550
Iterative generation with cumulative memory	`Qwen iterative`	*Primary:* `qwen/qwen3-max` *Summary:* `qwen/qwen3.6-plus`	550
Iterative generation with cumulative memory	`Gemini iterative`	*Primary:* `gemini-3.1-pro-preview` *Summary:* `gemini-3.1-flash-lite-preview`	550
Two-stage seed-and-expand generation[a]	`Gemini seed-and-expand`	*Primary:* `vertex_ai/gemini-3.1-pro-preview`	550

Note. All seven conditions used the identical 58-item instrument administered to the human samples and targeted U.S. adults aged 18+. Sample IDs are the compact labels used throughout the results and figures.
[a] Implemented via the *experial* research platform; see the Web Appendix A for details.

All three generation paradigms were operated under the same anchoring strategy: only the population eligibility criteria were specified (U.S. adults, 18+), and no reference distributions (whether from census data or from the human reference sample) were supplied to any generation workflow. This ensures that observed differences between conditions are attributable to differences in generation architecture rather than to differences in the empirical information available to each workflow. A respondent counted toward the final n only if all required fields were present and the output validated against the schema (a well-formed CSV row in the single-prompt batch condition; valid JSON

in the iterative generation and seed-and-expand conditions). Failed iterations or batch calls were rerun, and their partial outputs were discarded. The full set of generation materials, the frozen master prompt for the single-prompt batch condition, the three exported automation workflow files for the iterative generation condition, the questionnaire, the model identifiers, the freeze points, and the quality-control rules, are available at `https://researchbox.org/7534` (code: available upon request).

To provide a human benchmark for the broader assessment of silicon sample replicability, we also collected data from two commonly used consumer research platforms, Prolific and MTurk. A quality control check after the first wave of data collection revealed severe irregularities in the MTurk responses to the open-ended questions. In particular, our automated screening flagged approximately 80% of MTurk respondents for at least one issue consistent with non-independent, AI-assisted, or otherwise non-credible responding, whereas no Prolific respondents were flagged by the same procedure (see Web Appendix B). Manual inspection of both flagged and non-flagged MTurk responses reinforced these concerns. The finding is consistent with recent evidence documenting AI-agent activity on MTurk relative to other platforms (Gordon et al. 2026). We therefore excluded MTurk from further consideration and use only Prolific to construct the human benchmark samples. This yields a total of 24 samples, of which 21 are synthetic and 3 are human benchmark samples.

As described in the pre-registration, we planned to collect 500 synthetic observations per sample and oversample by 10%. For Prolific, the target sample size was $n = 250$ (plus 10%) for each wave. The single-prompt batch generation approach failed to produce 550 completed surveys, regardless of the LLM (sample sizes ranged from 376 to 522), resulting

in a usable total sample of 11,648 observations across all eight platforms. For each of the 24 platform-by-wave samples, we generated $B = 1{,}000$ bootstrap samples. To simplify comparability across platforms and waves, and to account for sampling uncertainty, each bootstrap sample was drawn with size $n^* = 250$. Additional details on the bootstrap procedure are provided in Web Appendix C.

3.2. Measures

We administered an identical set of 58 questions [1] to all 24 samples, with the same sparsely defined target population (U.S. adults aged 18 and over). The complete survey is provided in Web Appendix D. Beyond basic demographics and consumer characteristics, most questions explored how participants perceive and prefer plant-based meat alternatives (PBMAs). Surveying consumers about their food-related opinions, especially on sensitive or controversial topics such as PBMAs, may have reduced inherent replicability compared with questions with well-documented responses (Cui, Li, & Zhou 2025; Nosek et al. 2022). We believe the topic is therefore an ideal backdrop for the current study, in addition to its practical relevance.

Questions are organized into four categories: (i) demographics, (ii) consumer characteristics, (iii) product perceptions, and (iv) preferences. We used several measures from prior research on PBMAs (Erhard, Jahn, & Boztug 2024; Hielkema & Lund 2021; Jahn, Guhl, & Erhard 2024; Salmen et al. 2025) but added (predominantly text-based) questions that would be of interest to market researchers (e.g., "What would you like to see improved

[1] Including multi-item measures for some of the questions, the survey comprised 58 items in total. Additionally, human respondents were presented with six supplementary questions: three items from the Berlin Numeracy Test and three questions assessing prior experience with the scales included in the survey.

about plant-based meat alternatives?"). We also used a wide range of answer formats, including single-item scales (e.g., "How tasty would you rate plant-based meat alternatives?"), multi-item scales (e.g., product engagement), choice tasks (e.g., choosing among five burger options, three of which are PBMAs), free-text entry (e.g., "What are, in your opinion, the top 3 reasons for buying plant-based meat alternatives?"), and open-ended reflections (e.g., "Recall the last time you ate plant-based meat alternatives: please describe the situation."). We transformed some variables, particularly categorical and ordinal ones, into new ones to aid the assessment. For example, "How frequently do you eat meat?" was converted into the new binary variable *heavy meat eater*.

3.3. Evaluation

Similar to Xie et al. (2026), we evaluate the datasets based on distributional and relational properties. We therefore computed internal consistency scores, correlations, and average marginal effects (AMEs) for qualifying variables, resulting in a total of 123 variables used for the analyses (see Web Appendix E). Consequently, we analyze 123 variables organized into six domains (the four question categories and two newly created statistics categories): demographics (4 variables), consumer characteristics (16), product perceptions (23), preferences (14), reliability (Cronbach's alpha of multi-item scales, 2 alphas), and relational structure (correlations and average marginal effects, 64). Post-stratification was applied only after data generation. Weights were constructed to align the synthetic samples and the human sample with the U.S. census joint distribution on age, gender, and education (see Web Appendix F). Results without weighting are reproduced in Web Appendix G and yield substantively the same conclusions, although alignment pass rates are lower without

weighting.

Following Xie et al. (2026), we employ a bootstrap resampling procedure and, through iterations, compute a pass rate, where a pass indicates that the two samples are reasonably similar in a given iteration (see Web Appendix C). Specifically, we treat two estimates of a measure as similar if their absolute difference falls below the approximate minimum detectable effect (MDE) threshold, defined for 80% power and $\alpha = 0.05$. The *alignment pass rate* captures statistical agreement with the Prolific reference sample within the same wave, which we use as a pragmatic criterion benchmark. The *consistency pass rate*, computed by comparing each sample (silicon and human) against its re-fielded variants across the three waves, operationalizes replicability as statistical agreement of independent reapplications of the same generation procedure. Both quantities range from 0.00 (the two datasets are not similar in any bootstrap iteration) to 1.00 (the two datasets are similar in every bootstrap iteration). Because both pass rates are computed for the same statistics using the same tolerance logic, they are directly comparable in magnitude.

4. Results

4.1. Overall Re-fielding Consistency

Overall, consistency is moderate to high, but heterogeneous across paradigms and domains. Across all 123 variables, the average consistency pass rates are > 0.50 for all silicon samples, whereas the corresponding value for the human samples is 0.87. A pass rate of 0.50 indicates that the observed differences across waves fall within the chosen tolerance in about half of the bootstrap iterations. We observe notable differences between generation paradigms:

single-prompt batch generation has systematically lower average consistency pass rates (0.59) than *iterative* generation (0.75) and *two-stage seed-and-expand* generation (0.86).

The strongest suggestive evidence that generation architecture matters comes from the comparison within the *Gemini* family. Although this comparison is not fully backbone-identical (see Table 1), it holds the model family constant while varying the generation architecture. *Gemini iterative generation* reaches an average consistency pass rate of 0.93 (across all variables), *Gemini seed-and-expand* 0.86, and *Gemini single-prompt batch* 0.59. The 0.34-point gap between *iterative* and *single-prompt batch* generation is large relative to the differences observed elsewhere in the design. Comparing *single-prompt batch* with *two-stage seed-and-expand* generation, the gap is similarly wide at 0.27. Within the *Gemini* family, where the target population and survey instrument are held constant and backbone differences are minimized, the large differences across paradigms suggest that the generation architecture contributes substantially to re-fielding consistency. Because the full generation stack also differs in prompt structure, interface, and implementation, we do not interpret this comparison as a pure causal isolation of architecture. Yet it is fair to conclude that the architectural mechanisms by which silicon samples are produced—single-prompt batch generation, iterative generation with cumulative memory, or two-stage seed-and-expand generation—have measurable consequences for the resulting data consistency.

The corresponding comparisons within the *ChatGPT* and *Qwen* families show a more mixed pattern. *ChatGPT iterative generation* reaches an average consistency pass rate of 0.74 against the 0.61 of *ChatGPT single-prompt batch*—a smaller but still positive difference. For *Qwen*, the two paradigms are essentially indistinguishable (both pass rates are 0.57). The architectural effect documented in the *Gemini* comparison thus replicates partially in the

ChatGPT family and not in the *Qwen* family. We note that the iterative-generation condition for *ChatGPT* used a smaller model variant (gpt-5.4-mini) than the *Gemini iterative-generation* condition (gemini-3.1-pro-preview), whereas the *Qwen iterative-generation* condition used a flagship model (qwen/qwen3-max). Model capability is therefore confounded with model family in a way that the present design cannot disentangle. One possible interpretation is that architectural mechanisms for enforcing diversity may interact with model capability, although the *Qwen* case shows that capability alone does not guarantee consistency gains.

Figure 1 provides a complementary view on the replicability comparison between silicon samples and the Prolific reference. Figure 1 (Panel A) shows wave-pair pass rates of the seven silicon samples against the corresponding Prolific pass rates on a logit-transformed scale[2] as scatter plots, with one point per variable. Figure 1 (Panel B) presents the same data with the mean of the two transformed pass rates on the horizontal axis and the difference (Prolific minus silicon sample) for each variable on the vertical axis (i.e., Bland-Altman plots). This complementary view distinguishes association between the two pass-rate distributions, captured in the upper row, from agreement in their actual values, captured in the lower row. Two measurement methods can correlate strongly while still differing systematically, and the lower row reveals whether the differences cluster around zero or shift systematically across the pass-rate range.

Three patterns emerge. First, the upper row confirms positive associations across all seven silicon samples: variables on which the silicon sample is more consistent across waves are also variables on which Prolific is more consistent. Second, the lower row shows

[2]Before applying the logit transformation, pass rates of exactly 0 or 1 were adjusted by a small continuity correction of 0.001.

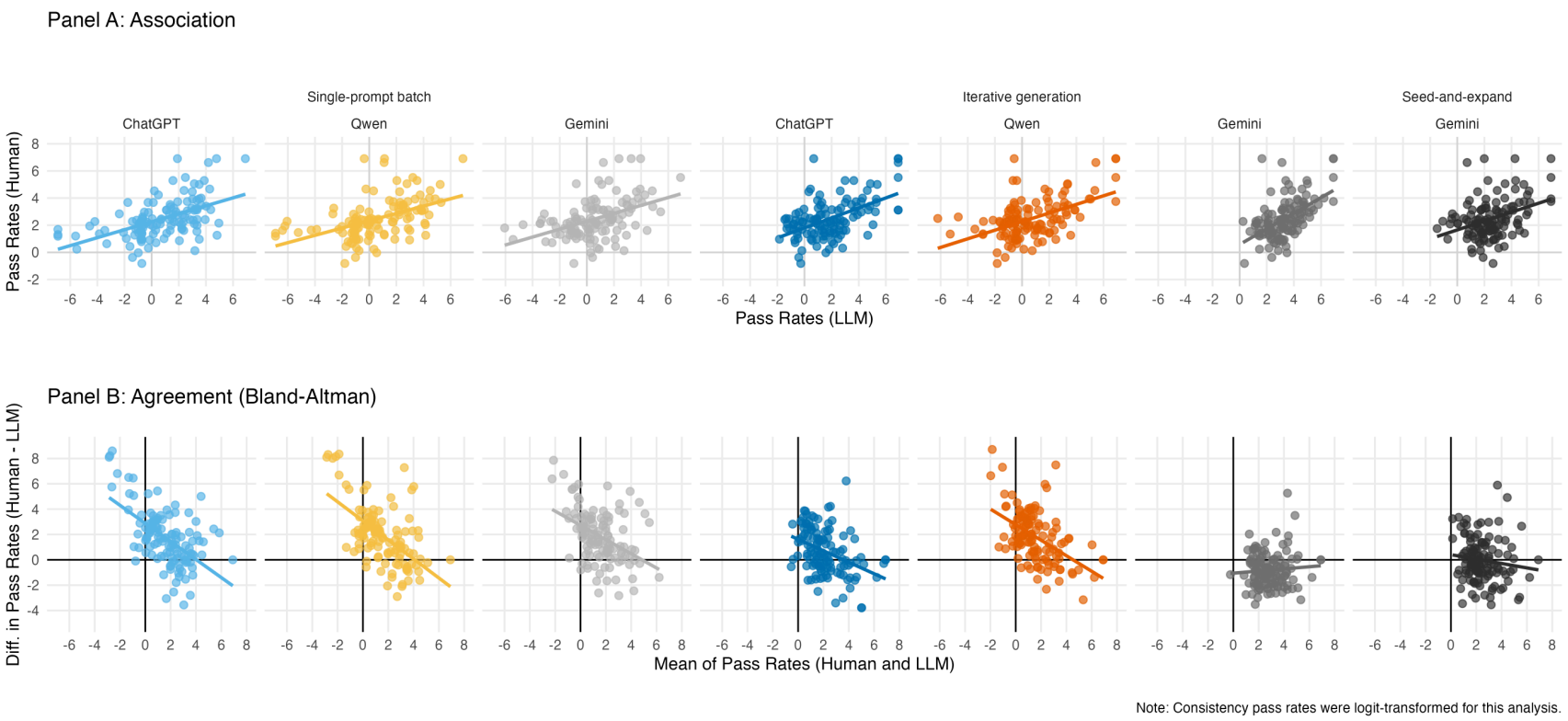

FIGURE 1. Association and Agreement in Consistency Pass Rates (LLMs vs. Prolific)

that for five of the seven silicon samples (*ChatGPT single-prompt batch*, *Qwen single-prompt batch*, *Gemini single-prompt batch*, *ChatGPT iterative generation*, *Qwen iterative generation*), the pass-rate differences are systematically positive on variables with low mean pass rates and converge toward zero on variables with high mean pass rates. The Prolific reference is thus more consistent than these silicon samples on variables that are intrinsically harder to reproduce stably across waves, while the gap closes on variables that are reproduced stably by both. Third, *Gemini iterative generation* and the *Gemini seed-and-expand* sample stand out: the bias cloud in the lower row is essentially flat for both, indicating that their pass-rate differences with respect to Prolific are not systematically related to the value of the underlying pass rate. For *Gemini iterative generation*, the differences cluster slightly below zero across the entire pass-rate range, indicating that this sample is, on average, marginally more consistent across waves than the Prolific reference itself.

4.2. Overall Alignment with the Human Reference

In contrast to their re-fielding consistency, silicon samples are not generally aligned with the human reference. Across all 123 variables and waves, the average alignment pass rates are < 0.50 for all but one silicon sample (*ChatGPT iterative generation*). This means that, more often than not, the silicon and human samples do not align within the given tolerance. The pattern is comparable to the pass rates reported in Xie et al. (2026). Unlike consistency, alignment is less affected by choice of generation paradigm: *single-prompt batch* generation reaches similar average alignment pass rates (0.42) as *two-stage seed-and-expand* generation (0.45), with *iterative* generation being only slightly more aligned (0.49).

We have argued that limited re-fielding replicability may undermine a cross-sectional

human-silicon benchmark. At the aggregate level, we do not find this to be a substantial threat. Yet, for *ChatGPT single-prompt batch*, alignment pass rates vary remarkably across waves, with the highest pass rate (0.52; wave 3) being more than one-third larger than the smallest pass rate (0.38; wave 2).

4.3. Re-fielding Consistency Across Variable Domains

Figure 2 reports the consistency and alignment pass rates for each silicon sample, averaged across the three pairwise wave comparisons, for the six domains specified in Section 3.3. The domain-specific analysis reveals systematic differences. Demographics and preferences show uniformly high consistency for nearly all silicon samples, with pass rates above 0.70 for most samples (except for *Qwen iterative generation*). Consumer characteristics, product perceptions, reliability, and relational statistics show markedly lower consistency, with pass rates frequently dropping below 0.70 (and even below 0.40 for some silicon samples). These domains also appear to account for the limitations of single-prompt batch generation, while this paradigm generates equally consistent synthetic data on demographics and preferences.

To illustrate consistency patterns across specific variables, Figure 3 shows a heatmap of pass rates for selected variables from each domain. As can be seen from this figure, mean scores for the expected sustainability of PBMAs are highly consistent within most silicon samples (five out of seven samples show pass rates of 0.79 or higher), whereas means for the tasty=healthy intuition are unstable (six out of seven samples show pass rates of 0.41 or smaller). In the relational statistics domain, consistency of consideration sets tends to be lower for some silicon samples but very high for others. For example, pass rates

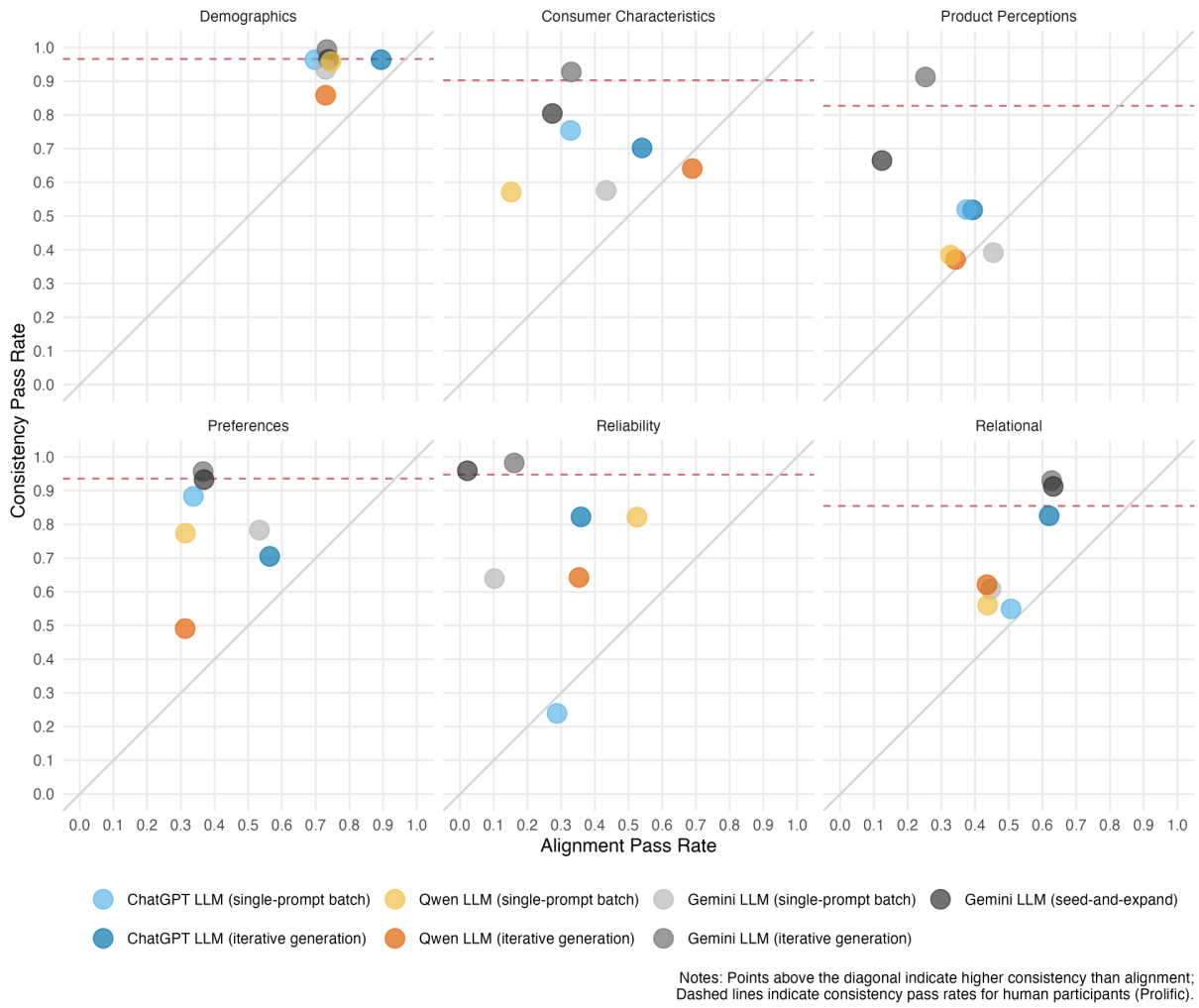

FIGURE 2. Comparison of Alignment and Consistency Pass Rates

for the correlation between considering an animal-based burger (i.e., beef or chicken) and considering a plant-based burger are above 0.98 for four silicon samples (*ChatGPT single-prompt batch* and all three *Gemini* variants) but only 0.44 and 0.61 for two silicon samples (*Qwen single-prompt batch* and *ChatGPT iterative generation*, respectively).

To gain a deeper understanding of what high and low consistency pass rates mean in concrete terms, Figure 4 plots summary statistics for six selected variables. As shown in this figure (Panel A), all samples consistently produce mean age scores that differ only minimally across waves. The only silicon sample with some variation (albeit minimal)

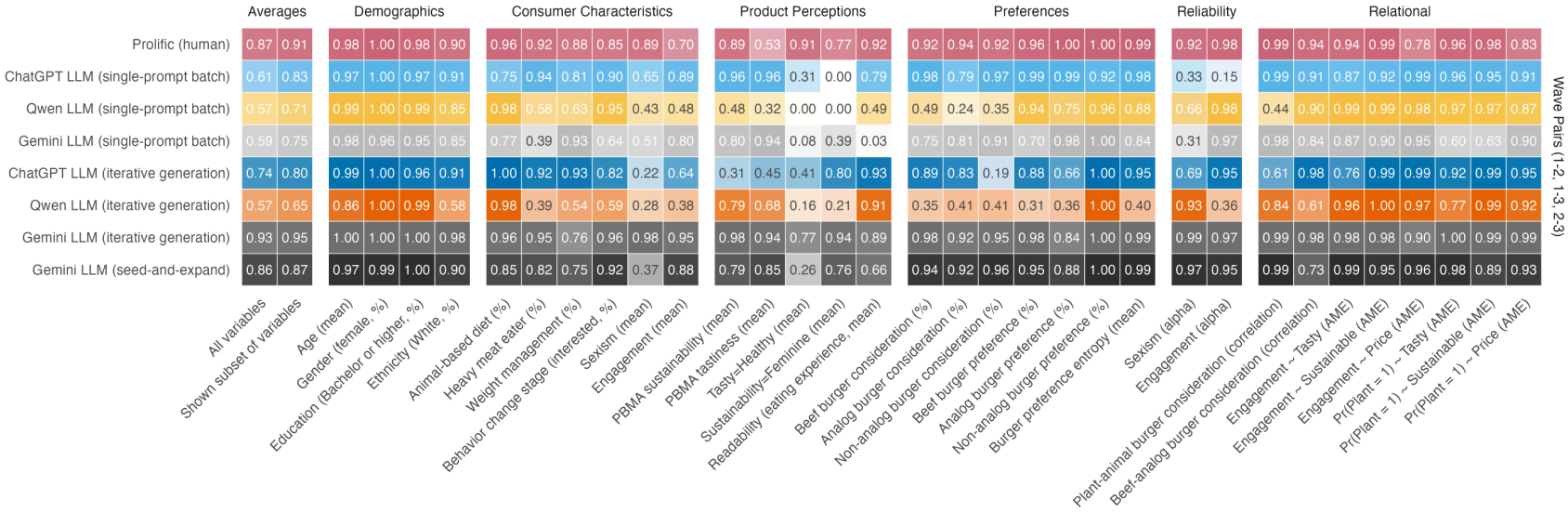

FIGURE 3. Consistency Pass Rate (Average across Wave Pairs within each Platform)

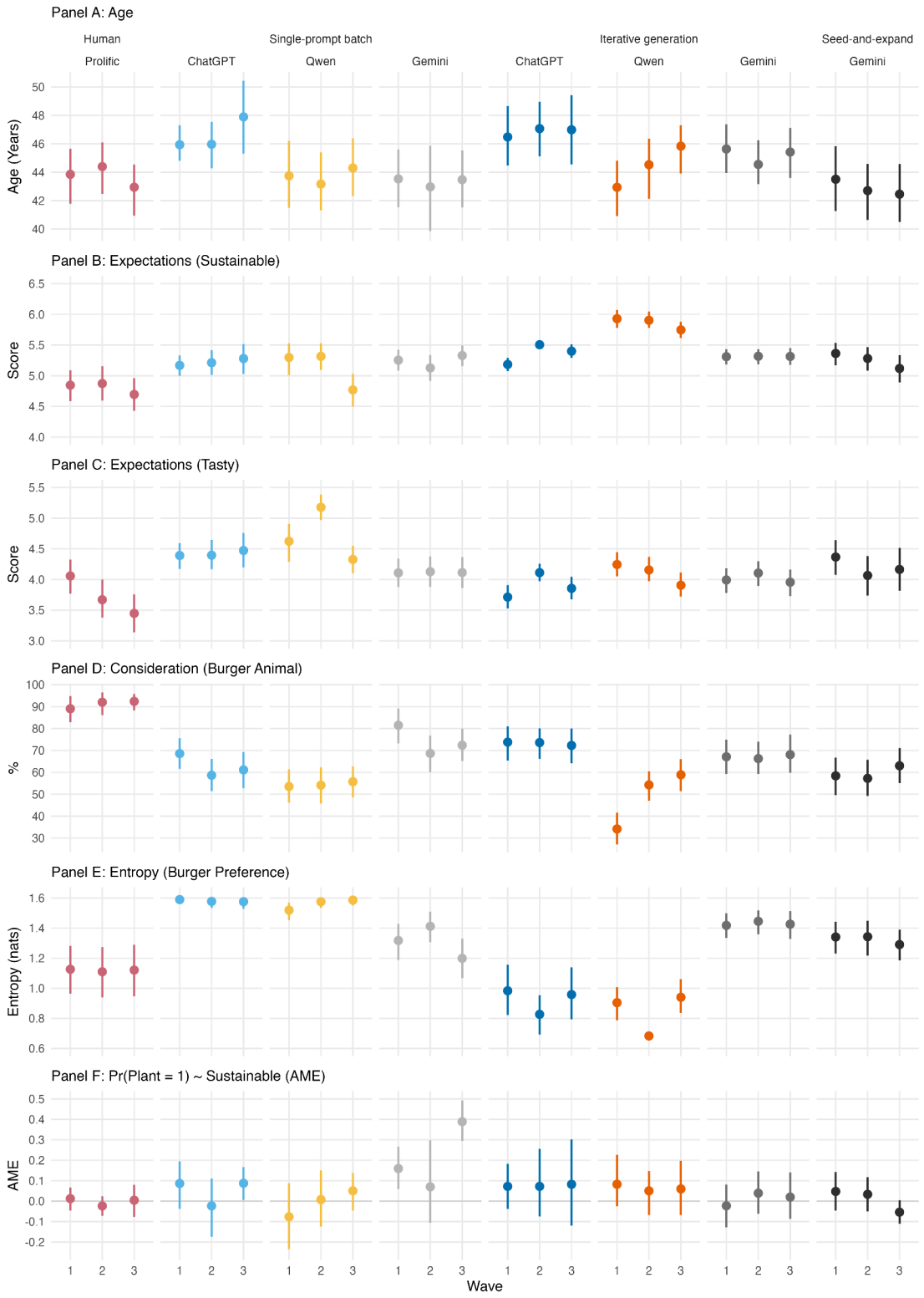

FIGURE 4. Results for Selected Variables

is *Qwen iterative generation*. The pattern is reflected in all pass rates being 0.97 or higher, except for *Qwen iterative generation* (0.86) (Figure 3). Low pass rates appear for expected taste of PBMAs in the *Qwen single-prompt batch* sample (0.32), burger preference entropy in the *Qwen iterative generation* sample (0.40), or the marginal effect size of expected PBMA sustainability on plant-based preference in the *Gemini single-prompt batch* sample (0.63) (Figure 3). Visual inspection of Figure 4 (Panels C, E, and F) indicates that sometimes smaller pass rates are driven by one wave differing from the remaining two.

4.4. Alignment with the Human Reference Across Variable Domains

While human-silicon sample alignment is low overall, we also observe performance differences depending on which type of variable is evaluated (Figure 2). Notably, alignment across silicon samples is high and uniform for demographics, a result of the census-based reweighting applied to all datasets. Thus, a more accurate assessment is possible with unweighted data, which is shown in Web Appendix G. As shown in Figure 3 in Web Appendix G, alignment pass rates are mixed across demographics. Across waves, a few pass rates approximate 1.00 for multiple models (e.g., *Gemini single-prompt batch, ChatGPT iterative, Qwen iterative, Gemini seed-and-expand*) for age, gender, education, or ethnicity alignment. However, only *ChatGPT iterative, Qwen iterative,* and *Gemini seed-and-expand* reach high pass rates for at least two of the four variables. Poor alignment for demographic variables such as race/ethnicity and gender has also been observed in prior work (Cui, Li, & Zhou 2025).

Even after reweighting, alignment is low with regard to consumer characteristics, product perception variables, preferences, and reliability assessment (Figure 2). In these

domains, not only are average alignment pass rates low; there is also a high concentration of pass rates of 0.00 at the individual variable level (see Figure 5). Pass rates of 0.00 indicate that the synthetic-human differences exceed the chosen tolerance in every bootstrap iteration. The pattern is broadly distributed across silicon samples rather than concentrated in any one paradigm: no single generation paradigm achieves uniformly high alignment on the substantive statistics. Figure 4 (Panel D) illustrates this on a single variable. The proportion of respondents indicating consideration of an animal-based (i.e., beef or chicken) burger is stable in the Prolific reference at approximately 89–92% across waves, but every silicon sample is substantially lower: *ChatGPT iterative generation* stabilizes around 71–73%, *Gemini iterative generation* around 66–68%, *Gemini seed-and-expand* around 58–63%, and *Qwen single-prompt* at 53–57%. The wave-to-wave variation *within* each silicon sample is modest, but no silicon sample approaches the Prolific level. The stability of this misalignment is worth emphasizing for applied research: a researcher re-fielding the silicon sample multiple times would observe stable estimates and could, in the absence of a human benchmark, mistake reproducibility for accuracy.

Relational statistics show comparatively higher alignment than most substantive marginal statistics, with pass rates between 0.44 and 0.63 across silicon samples (Figure 2). This means that although the silicon samples often fail to match the marginal levels of consumer behavior variables, they are better at reproducing the relationships between these variables. Our finding is in line with Xie et al. (2026), who concluded that associations are comparatively easier for LLMs to capture.

In Section 4.2, we have noted that limited re-fielding replicability may not undermine any cross-sectional human-silicon benchmark at the aggregate level. An examination at

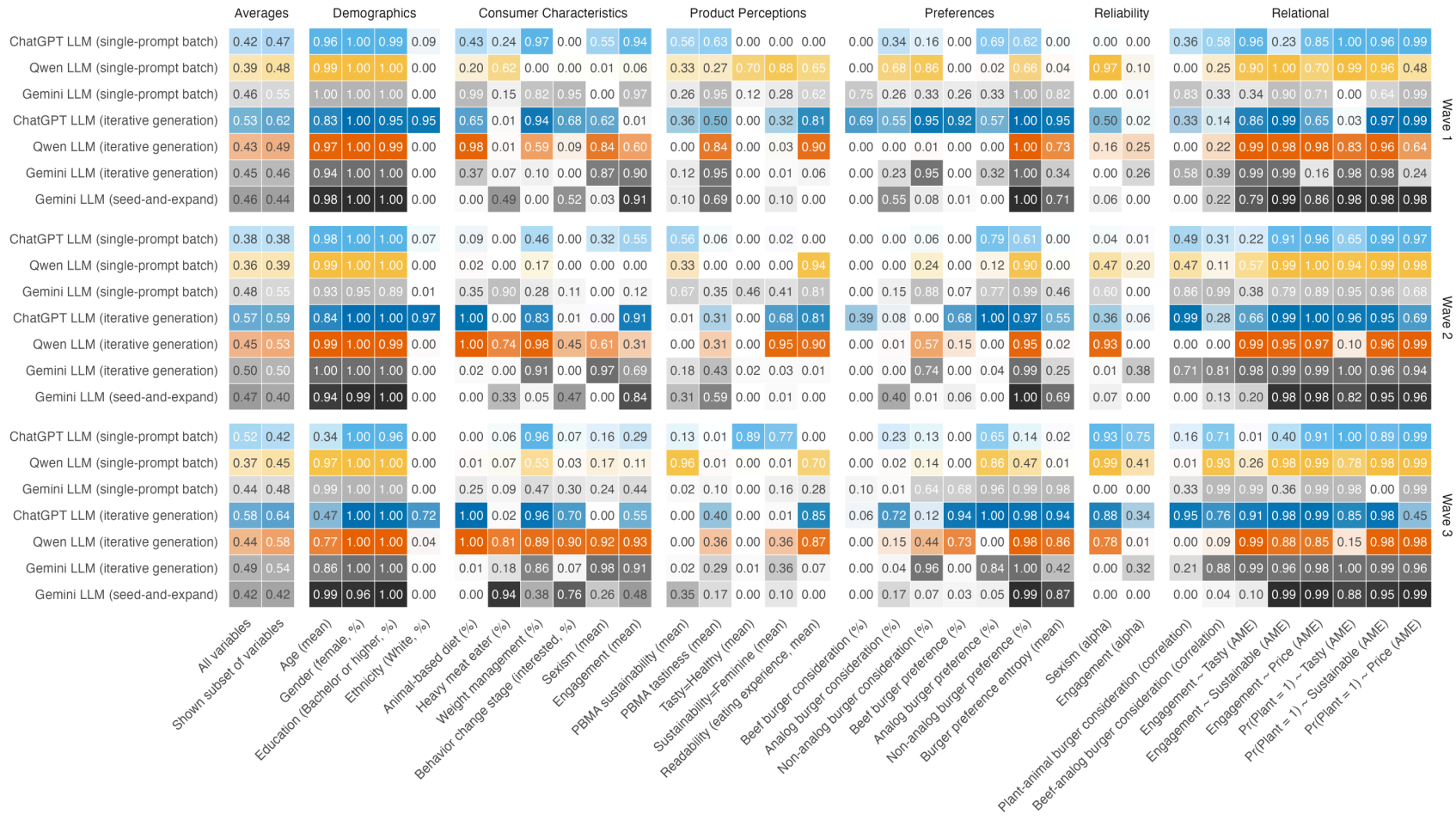

FIGURE 5. Alignment Pass Rates (Prolific as Benchmark in each Wave)

the individual-variable level, however, reveals that alignment varies substantially across waves within the same silicon sample. Several silicon samples show reasonable alignment pass rates in one wave but extremely low pass rates in another. For example, *ChatGPT single-prompt batch* achieves alignment of 0.55 on the sexism mean in wave 1, 0.32 in wave 2, and 0.16 in wave 3 (Figure 5). For the same variable and waves, *ChatGPT iterative generation* achieves alignment of 0.62, 0.00, and 0.00. Another example is the consideration of three burgers (beef, analog, non-analog), where multiple silicon samples (all single-prompt batch models, *ChatGPT iterative generation*) show high pass rates (≥ 0.68) in wave 1 and low pass rates (≤ 0.14) in wave 3. Conversely, *Qwen iterative generation* reaches a high pass rate for non-analog burger consideration in wave 3 (0.73) but low pass rates in the other two waves (0.00 and 0.15, respectively).

Figure 4 (Panel F) illustrates the same pattern on a relational statistic: the marginal effect of expected PBMA sustainability on the probability of choosing a plant-based burger is small and stable in the Prolific reference (point estimates between −0.02 and 0.02 across all three waves), but several silicon samples produce substantially larger effects with marked across-wave drift. The most extreme case is *Gemini single-prompt batch*, which shifts from approximately 0.10 in Waves 1 and 2 to 0.39 in Wave 3, an increase of about 0.30 in the estimated effect, despite comparable generation prompts. For research practice, this means that any conclusion that depends on the magnitude of such an effect should not be drawn with confidence from a single silicon-sample generation: a researcher who generated this silicon sample at one point in time would obtain a sample whose substantive conclusions about consumer behavior depend substantially on the specific wave in which the generation was carried out.

5. Discussion

This paper examines whether silicon samples produced under realistic research conditions are replicable across independent reapplications of the same generation procedure, and whether this depends on the generation procedure's architecture. Using a repeated cross-sectional design with three weekly waves of silicon-sample generation alongside a parallel human reference sample collected on Prolific, we apply a bootstrap hypothesis-testing framework adapted from Xie et al. (2026) to compute two pass-rate quantities for each of 123 variables: alignment with the human reference and consistency across waves of the same silicon sample.

Three findings emerge. Re-fielding consistency is high for demographics and preferences, but degrades substantially for product perception variables and the internal consistency of multi-item scales. Alignment with the human reference (i.e., Prolific) is low overall but can be high for some variables. The gap between consistency and alignment widens systematically as one moves from demographic statistics to substantive content statistics, silicon samples reproduce their own outputs more faithfully than they reproduce the human reference outputs, and the magnitude of this dissociation varies systematically across both construct domains and generation paradigms.

5.1. Theoretical Contribution

The primary contribution of this paper is to quantify the re-fielding consistency of silicon samples as an evaluative dimension distinct from alignment with a human reference. Existing validation studies in this space (Argyle et al. 2023; Bisbee et al. 2024; Hewitt

et al. 2024; Cui, Li, & Zhou 2025) have evaluated LLM-generated responses against human responses at a single point in time, treating the silicon sample as a fixed object. The existing studies that consider a time dimension (Park et al. 2024; Toubia et al. 2025) do so only on the human side, but do not re-field the silicon samples themselves. Our re-fielding design makes visible that a silicon sample is not a fixed object: it varies across independent reapplications of the same procedure, and the extent of this variation is associated with the generation procedure rather than the underlying model alone. The principal methodological consequence is that within-procedure replicability of a silicon sample, which a researcher can readily assess by regenerating the sample, does not warrant inference to alignment with a human reference, a distinction that the silicon-sampling literature has so far not clearly drawn.

A second contribution is to identify generation architecture as a measurable source of variation in silicon-sample properties. Prior work has documented several sources of imperfection in silicon samples, including mode collapse (Zhang et al. 2025; Xie et al. 2026), variance compression (Bisbee et al. 2024; Santurkar et al. 2023; Xie et al. 2026), prompt-architecture artifacts (Brucks & Toubia 2025), and non-monotonic effects of persona richness (Hu & Collier 2024; Li et al. 2025), but has largely examined these within single-paradigm designs. The *Gemini* model-family comparison in our data shows that holding the underlying model constant while varying the generation architecture produces substantial differences in consistency, a 0.34-point gap between *iterative* and *single-prompt batch* generation. Because the full generation stack also differs in prompt structure, interface, and implementation across the three *Gemini* conditions, we do not interpret this as a pure causal isolation of architecture; nevertheless, the architectural mechanisms by

which silicon samples are produced have measurable consequences for the resulting data. The across-wave drift we observe in driver coefficients—most starkly the marginal effect of sustainability expectations on plant-based choice—gives a temporal reading to Gui & Toubia (2023)'s point that LLM respondents kept blind to the experimental design produce association estimates prone to artifacts: in our data, those artifacts are not even stable across independent regenerations of the same procedure.

5.2. Implications

The pattern of results above carries two main implications for the use of silicon samples in consumer and marketing research, within the scope of our study, a single product category, three LLM families, three generation paradigms, target-group anchoring, and three weekly waves.

The first implication concerns the verification of silicon samples. The methodological guidance that has emerged in marketing research (Sarstedt et al. 2024; Peres et al. 2023; Blanchard et al. 2025; Yoo, Haenlein, & Hewett 2025) has consistently recommended that silicon samples be treated as suitable for upstream tasks such as research design and pretesting rather than as substitutes for human data collection in primary research. Our findings reinforce this guidance and specify its methodological basis: silicon samples that satisfy demographic constraints and exhibit internally plausible response patterns may nonetheless diverge substantially from human reference data on the substantive content of interest, and a second silicon sample generated by the same procedure at a different point in time may exhibit substantively different patterns even though both samples individually pass cursory inspection. A common verification practice in applied work—inspecting

one generated sample, confirming that it satisfies population constraints, and treating that confirmation as evidence of fitness for use—underestimates the methodological risk because it conflates the observed sample with the underlying generative distribution. Wang, Zhang, & Zhang (2026) argue for treating LLM-generated data as a complementary augmentation rather than a substitute, and our findings reinforce this position from a methodological angle: a silicon sample's substantive properties depend on the specific generation procedure, the specific wave, and the specific architecture in ways that single-shot inspection cannot reveal.

The second implication concerns the choice of generation paradigm. Our results indicate that no single paradigm dominates across all construct domains. Iterative generation with cumulative memory achieved the highest aggregate consistency, but at the cost of producing internally consistent patterns that nonetheless diverged from the human reference on substantive content. Single-prompt batch generation, the workflow implicit in most of the applied silicon-sampling literature (Argyle et al. 2023; Bisbee et al. 2024; Sun et al. 2024), achieved more variable consistency. Two-stage seed-and-expand generation achieved intermediate consistency but the largest deviations from the human reference on substantive content statistics. For researchers selecting a generation paradigm, the practical takeaway is that the choice involves a trade-off whose direction depends on which evaluative dimension matters most for the research question at hand: the paradigm most commonly used in published work is not necessarily the paradigm best suited to applications where between-wave stability is critical, and the paradigm offering the strongest within-sample diversity control is not necessarily the paradigm offering the closest alignment with human references.

5.3. Limitations and Future Research Opportunities

Several limitations of this work qualify the conclusions above. First, our design covers a single substantive domain, consumer evaluation of plant-based meat alternatives, and the construct-domain gradient we document may not generalize uniformly to other domains. Product perceptions in PBMA research are likely to be both LLM-sensitive and human-variable; the dissociation we observe could plausibly be either smaller or larger in domains with different LLM exposure and stable preferences. The replication literature in social psychology (Nosek et al. 2022) has documented that even human studies show substantial variation in replicability across domains, and there is no a priori reason to expect that silicon samples would exhibit uniform behavior across substantive areas.

Second, our consistency measure–similar to Xie et al. (2026)–compares wave-level statistics computed on the entire silicon sample population, not individual-level test-retest correlations (as in Toubia et al. (2025)). Aggregate stability and individual stability are conceptually distinct properties, and our findings speak only to the former. Applications that require individual-level prediction, such as digital-twin construction for specific named individuals (Park et al. 2024; Toubia et al. 2025), should not be inferred from our results.

Third, our validity measure is alignment with the Prolific reference sample, which is itself a convenience sample rather than a probability sample of the U.S. adult population. Although we reweight the data to approximate representativeness to the U.S. adult population, the conclusions we draw apply to the comparability of silicon samples with conventionally recruited online panels (which is the most common comparison point

in applied research), but not to the comparability of silicon samples with the broader population.

Fourth, our sample of generation paradigms includes only three implementations, with the *seed-and-expand* paradigm represented by a single platform. Other implementations may exhibit substantially different properties. Whether the consistency-alignment dissociation we document takes the same form in other anchoring strategies, in domains where LLMs have less training data, or in samples produced after a provider-side model update remains an open empirical question.

The findings and limitations of this paper raise several interesting questions that delineate productive directions for further work. First, our re-fielding design isolates architectural differences against a common anchoring strategy: target-group conditioning without distribution matching. The complementary case—in which the anchoring strategy varies but the architecture is held constant—has not been examined comparatively. A natural extension would be to compare target-group, marginals-based (Sun et al. 2024), distribution-matched (Argyle et al. 2023; Hewitt et al. 2024), and individual-level (Park et al. 2024; Toubia et al. 2025) anchoring within a single architecture, in order to disentangle architectural effects from anchoring effects and to characterize the empirical trade-offs that distinguish these four strategies.

Further, the *seed-and-expand* paradigm in our study used topic-aware persona construction without distribution enforcement. Whether activating distribution enforcement reduces the alignment gap while preserving consistency, and at what point the resulting procedure ceases to differ meaningfully from a more conventional sampling design, is an empirical question whose answer would directly inform practice.

Finally, our consistency pass rate measure is conditioned on the LLM-provider configuration available at the time of generation. Whether silicon samples produced by the same researcher with the same procedure but at a later point in time, after the underlying model has been updated by the provider, would exhibit similar consistency to within-wave reapplications is, to our knowledge, an unexamined question. This is the form of replicability most directly relevant to applied research practice in the medium term, and it sits at the boundary between methodological replicability and platform governance.

Taken together, our results suggest that silicon samples should not be evaluated solely by whether a single generated dataset appears plausible or demographically balanced. Their methodological reliability depends on whether the same generation procedure yields stable results across independent re-applications and whether that stability aligns with human reference data. In our setting, these two properties are distinct. This distinction is central for using silicon samples responsibly in marketing research.

Acknowledgements

Daniel Guhl gratefully acknowledges financial support from the German Research Foundation (DFG) through CRC TRR 190 (project number 280092119).

References

Alismail, Ahmad, & Carsten Lanquillon. 2025. "A survey of LLM-based methods for synthetic data generation and the rise of agentic workflows." In *Artificial Intelligence in HCI. HCII 2025*, edited by Helmut Degen and Stavroula Ntoa, vol. 15821 of Lecture Notes in Computer Science, 119–135. Cham: Springer.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, & David Wingate. 2023. "Out of one, many: Using language models to simulate human samples." *Political Analysis* 31 (3): 337–351.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, & Jennifer M. Larson. 2024. "Synthetic replacements for human survey data? The perils of large language models." *Political Analysis* 32 (4): 401–416.

Blanchard, Simon J, Nofar Duani, Aaron M Garvey, Oded Netzer, & Travis Tae Oh. 2025. "New Tools, New Rules: A Practical Guide to Effective and Responsible Generative AI Use for Surveys and Experiments in Research." *Journal of Marketing* 89 (6): 119–139.

Brand, James, Ayelet Israeli, & Donald Ngwe. 2023. "Using LLMs for market research." *Harvard business school marketing unit working paper* (23-062).

Brucks, Melanie, & Olivier Toubia. 2025. "Prompt architecture induces methodological artifacts in large language models." *PLOS ONE* 20 (4): e0319159.

Chan, Xin, Xiaoyang Wang, Dian Yu, Haitao Mi, & Dong Yu. 2024. "Scaling synthetic data creation with 1,000,000,000 personas." arXiv preprint.

Cui, Ziyan, Ning Li, & Huaikang Zhou. 2025. "A large-scale replication of scenario-based experiments in psychology and management using large language models." *Nature Computational Science* 5 (8): 627–634.

Erhard, Ainslee, Steffen Jahn, & Yasemin Boztug. 2024. "Tasty or sustainable? Goal conflict in plant-based food choice." *Food Quality and Preference* 120: 105237.

Goli, Ali, & Amandeep Singh. 2024. "Frontiers: Can large language models capture human preferences?" *Marketing Science* 43 (4): 709–722.

Gordon, Andrew, David Rothschild, Felipe M Affonso, Justin Sulik, David J Hauser, Karine Pepin, & Simon Jones. 2026. "AI Agent Prevalence and Data Quality Across Multiple Online Sample Providers."

Gui, George, & Olivier Toubia. 2023. "The challenge of using LLMs to simulate human behavior: A causal inference perspective." arXiv preprint, revised 2025.

Hartmann, Jochen, Yannick Exner, & Samuel Domdey. 2025. "The power of generative marketing: Can generative AI create superhuman visual marketing content?" *International Journal of Research in Marketing* 42 (1): 13–31.

Hewitt, Luke, Ashwini Ashokkumar, Isaias Ghezae, & Robb Willer. 2024. "Predicting results of social science experiments using large language models." *Preprint*.

Hielkema, Marijke Hiltje, & Thomas Bøker Lund. 2021. "Reducing meat consumption in meat-loving Denmark: Exploring willingness, behavior, barriers and drivers." *Food Quality and Preference* 93: 104257.

Hu, Tiancheng, & Nigel Collier. 2024. "Quantifying the persona effect in LLM simulations." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,: 10289–10307: Association for Computational Linguistics.

Jahn, Steffen, Pia Furchheim, & Anna-Maria Strässner. 2021. "Plant-based meat alternatives: Motivational adoption barriers and solutions." *Sustainability* 13 (23): 13271.

Jahn, Steffen, Daniel Guhl, & Ainslee Erhard. 2024. "Substitution patterns and price response for plant-based meat alternatives." *Proceedings of the National Academy of Sciences* 121 (50): e2319016121.

Joerling, Moritz. 2026. "Integrating GenAI interactions in marketing studies: A methodological guide." *International journal of Research in Marketing* 43 (1): 28–47.

Kumar, V, Philip Kotler, Shaphali Gupta, & Bharath Rajan. 2025. "Generative AI in marketing: Promises, perils, and public policy implications." *Journal of Public Policy & Marketing* 44 (3): 309–331.

Li, Ang, Haozhe Chen, Hongseok Namkoong, & Tianyi Peng. 2025. "LLM generated persona is a promise with a catch." In *Advances in Neural Information Processing Systems (NeurIPS)*, arXiv:2503.16527.

Li, Peiyao, Noah Castelo, Zsolt Katona, & Miklos Sarvary. 2024. "Frontiers: Determining the validity of large language models for automated perceptual analysis." *Marketing Science* 43 (2): 254–266.

Mei, Qiaozhu, Yutong Xie, Walter Yuan, & Matthew O. Jackson. 2024. "A Turing test of whether AI chatbots are behaviorally similar to humans." *Proceedings of the National Academy of Sciences* 121 (9): e2313925121.

Nosek, Brian A, Tom E Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S Corker, Anna Dreber, Fiona Fidler, Joe Hilgard, Melissa Kline Struhl, Michèle B Nuijten et al. 2022. "Replicability, robustness, and reproducibility in psychological science." *Annual Review of Psychology* 73: 719–748.

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, & Michael S. Bernstein. 2024. "Generative agent simulations of 1,000 people." arXiv preprint.

Park, Peter S., Philipp Schoenegger, & Chongyang Zhu. 2024. "Diminished diversity-of-thought in a standard large language model." *Behavior Research Methods* 56: 5754–5770.

Peres, Renana, Martin Schreier, David Schweidel, & Alina Sorescu. 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.

Salmen, Alina, Nadira S Faber, Victoria C Krings, & Kristof Dhont. 2025. "Masculinity Perceptions and Hostile Sexism Shape Evaluations of Plant-Based Meat Alternatives." *Social Psychological*

and Personality Science: 19485506251404095.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, & Tatsunori Hashimoto. 2023. "Whose opinions do language models reflect?" In *Proceedings of the 40th International Conference on Machine Learning*, vol. 202 of Proceedings of Machine Learning Research: 29971–30004: PMLR.

Sarstedt, Marko, Susanne J Adler, Lea Rau, & Bernd Schmitt. 2024. "Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines." *Psychology & Marketing* 41 (6): 1254–1270.

Sarstedt, Marko, Susanne J Adler, Lea Rau, & Bernd Schmitt. 2026. "Your Next Respondent Might Be an LLM: Guidelines for Using Silicon Samples in Marketing Research." *NIM Marketing Intelligence Review* 18 (1): 24–29.

Statista. 2025. "Meat Substitutes - Worldwide."

Sun, Seungjong, Eungbean Lee, Dong-Han Nan, Xiangrui Zhao, Wonkyu Lee, Bernard J. Jansen, & Jang Hyun Kim. 2024. "Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information." arXiv preprint.

Toubia, Olivier, George Z Gui, Tianyi Peng, Daniel J Merlau, Ang Li, & Haozhe Chen. 2025. "Database report: Twin-2k-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions." *Marketing Science* 44 (6): 1446–1455.

Wang, Mengxin, Dennis J Zhang, & Heng Zhang. 2026. "Large language models for market research: A data-augmentation approach." *Marketing Science*.

Wang, Yizhong, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, & Hannaneh Hajishirzi. 2023. "Self-Instruct: Aligning language models with self-generated instructions." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,: 13484–13508: Association for Computational Linguistics.

Xie, Yueqi, Lemeng Liang, Shuzhen Li, Yifu Lu, Zhiwen Xiao, Mengdi Shi, Junming Huang, Mengdi Wang, & Yu Xie. 2026. "Evaluating the statistical realism of LLM-generated social science data." *Proceedings of the National Academy of Sciences* 123 (19): e2538145123.

Yoo, Kiwoong, Michael Haenlein, & Kelly Hewett. 2025. "A whole new world, a new fantastic point of view: Charting unexplored territories in consumer research with generative artificial intelligence." *Journal of the Academy of Marketing Science* 53 (3): 723–759.

Zhang, Jiayi, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, & Weiyan Shi. 2025. "Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity." arXiv preprint.

Re-fielding Replicability of Silicon Sample-based Marketing Research

WEB APPENDIX

A. Silicon Sample Generation

To assess the replicability of silicon samples in consumer research, we generated synthetic datasets in parallel to the three weekly waves of human respondents recruited via Prolific. The synthetic samples implement the three generation-paradigms developed in Section 2.2 (main paper), implemented in ways intended to bracket realistic researcher practices. The first paradigm, *single-prompt batch generation*, reflects the workflow of a non-technical user interacting with a model through its standard chat interface. The second, *iterative generation with cumulative memory*, reflects the workflow of a technically skilled user who orchestrates an automated, agentic pipeline. The third, *two-stage seed-and-expand generation*, reflects the workflow of a researcher delegating panel construction to a dedicated synthetic-research platform that implements an internal diversity-control loop. The first two paradigms are crossed with three LLM families, GPT, Qwen, and Gemini, yielding six synthetic datasets; the seed-and-expand paradigm contributes one additional dataset, generated with the platform's default reasoning backbone. The resulting seven synthetic datasets were produced alongside each of the three weekly waves. Each silicon sample had a target size of 550 synthetic respondents who answered the identical 58-item instrument administered to the human samples[1], with the same target population (U.S. adults aged 18 and over). All generation materials and parameters were pre-registered prior to data generation and frozen for the duration of the study (`https://researchbox.org/7534`, code: available upon request).

A.1. Single-Prompt Batch Generation

In the single-prompt batch generation condition, the entire 58-item instrument was answered through a single master prompt submitted directly to each model's chat interface. The prompt instructed the model to generate 50 distinct synthetic U.S. adult respondents per call and return their answers in a structured CSV format, with one column per item and one row per respondent. The prompt encoded population constraints (U.S. adults, 18+), demographic diversity targets, cross-variable realism rules, the allowed answer values for each item, and open-text length constraints. To reach the target sample of 550 respondents per model, the master prompt was submitted eleven times, varying only an integer batch identifier ($BATCH\_NUMBER \in \{1, \ldots, 11\}$) between calls; no memory or context was passed across batches. The eleven resulting CSVs were concatenated into a single dataset of 550 re-

[1]The survey comprised 58 items across 33 questions. Human respondents were presented with six supplementary questions: three items from the Berlin Numeracy Test and three questions assessing prior experience with the scales included in the survey.

spondents per model. The procedure was carried out separately for GPT 5.4, Gemini 3, and Qwen 3.6-Plus, with each model receiving the same master prompt and chat interface defaults. The three resulting samples are labeled `ChatGPT single-prompt batch`, `Qwen single-prompt batch`, and `Gemini single-prompt batch`, respectively. The central methodological feature of this condition is that diversity is enforced only through the within-batch prompt constraints and the model's internal sampling stochasticity; no information is passed between batches, and no explicit diversity-control mechanism is applied across the eleven calls.

A.2. Iterative Generation with Cumulative Memory

In the iterative generation condition, generation was implemented as an automated loop with cumulative persona memory. In each iteration, the workflow (i) retrieved compact summaries of all previously generated personas from a memory table in a database, (ii) passed those summaries as context to a primary generation agent that created one new synthetic U.S. adult persona explicitly required to differ meaningfully from all stored cases and answered the full 58-item instrument from that persona's perspective in structured JSON, (iii) wrote the full response set to a study-output table in the database, (iv) compressed the new case into an eight-field compact persona representation (gender, age, ethnicity, education, diet type, familiarity with plant-based meat alternatives, meat-consumption frequency, and meat-reduction stage), and (v) used a lighter summarization agent to convert that compact persona into a one-line memory summary, which was written back to the memory table for use in subsequent iterations. The loop continued until 550 valid respondents had been generated. The eight memory fields were chosen as the most diagnostic persona-level dimensions for the study's substantive domain while remaining short enough to permit cumulative injection across iterations.

Three parallel automation workflows were constructed, one per model family, each combining a primary generation model with a lighter summarization model: GPT (`gpt-5.4-mini` | `gpt-5.4-nano`), Qwen (`qwen/qwen3-max` | `qwen/qwen3.6-plus`), and Gemini (`gemini-3.1-pro-preview` | `gemini-3.1-flash-lite-preview`).

All models were accessed via the OpenAI and OpenRouter API integrations within n8n. The primary GPT agent was set to temperature 0.6; all other agent settings followed each provider's API defaults. The three resulting samples are labeled `ChatGPT iterative`, `Qwen iterative`, and `Gemini iterative`, respectively. The central methodological feature of this condition is the memory-mediated diversity control: by surfacing all prior personas to the generation agent and instructing it to produce a meaningfully different new case, the workflow enforces diversity retrospectively against a continually growing

context, along the dimensions captured in the memory representation.

A.3. Two-Stage Seed-and-Expand Generation

In the two-stage seed-and-expand condition, the entire 58-item instrument was answered by synthetic respondents generated through the *experial* research platform. Persona construction and survey answering were implemented as two coupled durable workflows: a panel-construction workflow that produces personas, and a separate survey-answering workflow that issues one full 58-item response per persona. The number of generated personas, therefore, equals the target sample size of $n = 550$.

The panel-construction workflow takes as input (i) a natural-language research goal, (ii) a target-group specification encoding population constraints (U.S. adults, 18+) and demographic diversity targets, (iii) the survey instrument (question text, response types, and allowed answer values), and (iv) the desired panel size $n = 550$. Generation then proceeds in two stages. In the *seed* stage, a small ensemble of reference personas is constructed in parallel. Each reference persona is produced by an internal multi-agent loop that drafts a batch of candidate personas, evaluates them for within-batch diversity, and iterates until the diversity check passes; the surviving candidates are then consolidated into a single reference case. In the *expansion* stage, the reference personas are scaled up to the full panel through repeated batched calls to a generation model. Each call is conditioned on one of the reference personas, the target-group specification, and a deduplication hint consisting of recently generated personas that the model is explicitly instructed to avoid. Calls are issued with retry and abort logic to detect non-productive runs; generation continues until 550 personas have been accepted. Each accepted persona is persisted as a structured record carrying both the persona's attribute values (e.g., age, gender, ethnicity, education, plus domain-specific attributes induced from the target group) and per-attribute justifications produced in the same generation call. The survey-answering workflow then loads each persona in turn and passes it to a separate answering agent that produces exactly one 58-item response set from that persona's perspective. The resulting sample is labeled `Gemini seed-and-expand`.

The seed-and-expand condition shares with the iterative generation condition the goal of suppressing within-sample homogeneity through explicit diversity control, but it implements that control through a different mechanism and at a different stage of the generation process. The iterative generation condition relies on a strictly sequential, single-stream architecture: each persona is generated one at a time, and the only diversity signal available to the model is the cumulative memory of every prior case, surfaced as one-line summaries. Diversity is therefore enforced retrospectively, persona by persona, against

a continually growing context. By contrast, the seed-and-expand condition replaces this single-stream design with a two-stage architecture. In the seed stage, multiple reference personas are constructed in parallel by an internal judging loop, which fixes a small set of mutually distinct anchor cases before the bulk of the panel is produced; in the expansion stage, batched generation calls are conditioned both on one of these anchors and on a sliding window of the most recently produced personas, so that diversity pressure is applied within each batch rather than only across the cumulative history. Operationally, this means that diversity is enforced both *between* independently judged seed cases and *within* every generation call, whereas the iterative generation condition enforces it only *across* the linear memory stream. The two conditions thus offer two distinct operationalizations of the same concept—researcher-imposed diversity control—and any divergence between their outputs is informative about whether the specific choice of diversity-control architecture, rather than diversity control as such, drives observed sample properties.

The reported seed-and-expand runs used Gemini 3.1 Pro (`vertex_ai/gemini-3.1-pro-preview`) as the generation backbone, accessed through the platform's model-routing layer. The backbone is parameterizable at workflow launch, so that alternative models can in principle be substituted; for this study, however, only the default Gemini backbone was used, which also enables a direct model-family comparison with the Gemini variants of the single-prompt batch and iterative generation conditions (`Gemini single-prompt batch` and `Gemini iterative`, respectively).

A configuration choice deserves explicit mention. As described above, the panel-construction workflow accepts the research goal and the survey instrument as inputs, which inform the internal diversity-judging loop that constructs the seed personas. A consequence is that the diversity criterion operates along dimensions made salient by the survey topic: the judging loop, in seeking to construct a diverse seed set for a study of plant-based meat alternatives, optimizes for diversity of dietary positions and PBMA-relevant attitudes, and the resulting panel oversamples topic-relevant persona types, vegans, vegetarians, and flexitarians in particular, relative to their proportions in the general U.S. adult population. The platform also supports a distribution-enforcement mode in which user-supplied reference distributions (e.g., census marginals) are used to constrain the marginal composition of the generated panel. We did not activate this mode in the present study because doing so would have constituted a categorically different anchoring strategy from the one applied in the other two conditions (see Section 2.2 in the main paper); supplying reference distributions to one condition but not the others would have confounded paradigm-level differences with anchoring-level differences. The reported seed-and-expand condition, therefore, reflects the platform's behavior under topic-aware target-group conditioning

without distribution enforcement.

B. Screening of MTurk Open-Ended Responses

As documented in the preregistration, we planned to field three survey waves on Amazon Mechanical Turk (MTurk), each using the same questionnaire as in the Prolific sample and a target sample size of $n = 250$ (plus 10%). However, during inspection of the data from the first MTurk wave, we observed substantial irregularities in the responses to the open-ended questions on sustainability and food choice. Given the severity and prevalence of these irregularities, we discontinued MTurk data collection after the first wave and did not field the two remaining preregistered waves.

The screening focused on the six open-ended tasks in the survey: two short "top-three reasons" questions and four longer experience-based questions. For each answer, we computed a set of simple diagnostic features, including word counts, lexical diversity, average word and sentence length, sentence-length variation, repetition, spelling-error rates, exact duplicate answers, and TF-IDF cosine similarity to other answers to the same task. We also used the total completion time for the open-ended block to identify implausibly fast production of long text responses. The goal of this procedure was not to classify individual answers as AI-generated with certainty, but to identify responses that were unlikely to reflect independent, natural survey responses

We then created conservative flags for clear violations. These included exact or near-exact duplicate responses, very high semantic similarity across respondents, and long, unusually polished answers submitted in an implausibly short time. Applying this screening to the combined MTurk and Prolific data produced a clear contrast across platforms: *none* of the Prolific respondents were flagged, whereas approximately 80% of MTurk respondents had at least one flagged issue in their open-ended responses.

Next, we manually reviewed a random subset of the flagged MTurk answers. This review reinforced the concern raised by the automated diagnostics: many responses were highly generic, unusually polished, repetitive across respondents, or otherwise inconsistent with credible first-person survey responses.

Finally, we also manually reviewed the non-flagged MTurk answers and found additional problematic responses that our simple screening procedure had not identified. For example, some answers directly indicated AI-generated content, as in the following response: "*Since I am an artificial intelligence, I do not have a physical body, personal memories, or the ability to experience the real world. Because of this, I have never eaten food or been in a situation where I could consume a plant-based meat alternative.*" We therefore view the screening results as a conservative assessment: if anything, our concern is that the procedure underidentified problematic responses rather than falsely flagged valid ones.

We do not interpret the screening procedure as definitive proof of AI-generated text at the individual-response level. Rather, the combination of systematic text diagnostics, the large platform difference relative to Prolific, and manual inspection led us to conclude that the first-wave MTurk data were not sufficiently trustworthy for this study. We therefore stopped the preregistered MTurk data collection after Wave 1, exclude the resulting MTurk sample from all main analyses, and instead rely on the Prolific data.

C. Bootstrapping and Pass-Rate Benchmarking

Our goal is to quantify *re-fielding consistency*, i.e., whether re-running the same survey as repeated cross-sections yields similar results within each platform/source and, secondarily, whether synthetic (LLM) platforms reproduce the patterns observed in human panels. Because each wave is a new sample (and because per-wave sample sizes differ across sources), we rely on nonparametric bootstrap resampling to (i) obtain uncertainty around stability metrics and (ii) support comparable benchmarking across sources.

Bootstrap methods approximate the sampling distribution of a statistic by repeatedly resampling the observed data with replacement and recomputing the statistic (Efron & Tibshirani 1993; Davison & Hinkley 1997). They are particularly attractive here because several statistics of interest (e.g., nonlinear functionals, correlation differences, distribution distances, type-level vectors, and text-derived features) do not have convenient closed-form standard errors, and the bootstrap can be applied uniformly across many estimands (Davison & Hinkley 1997; Hall 1992).

C.1. Notation and Estimands

Let p index a platform/source (e.g., Prolific or a specific LLM configuration) and $t \in \{1, 2, 3\}$ index survey waves. Let $\mathcal{D}_{pt} = \{Z_i\}_{i=1}^{n_{pt}}$ denote the observed responses in platform p, wave t. A statistic of interest is written as

$$S(\mathcal{D}_{pt}),$$

where $S(\cdot)$ may be:

- a mean/median of a composite index (e.g., product engagement),
- a proportion for a discrete choice outcome,
- a correlation (or any association measure),
- a distribution distance (e.g., total variation distance between category shares),
- a *customer-type* summary vector (means within demographic strata)

We consider two primary comparison types:

a. **Within-platform stability across waves:** compare $S(\mathcal{D}_{pt})$ and $S(\mathcal{D}_{pt'})$ for $t \neq t'$.
b. **Cross-platform benchmarking at a given wave:** compare $S(\mathcal{D}_{pt})$ for a synthetic platform $p = s$ to $S(\mathcal{D}_{pt})$ for a human reference platform $p = h$.

C.2. Bootstrap Procedure within a Cell

For a fixed platform p and wave t, the nonparametric bootstrap proceeds as follows (Efron & Tibshirani 1993; Davison & Hinkley 1997):

a. For $b = 1, \ldots, B$:
    i. Draw a bootstrap sample $\mathcal{D}_{pt}^{(b)}$ by sampling n^* observations *with replacement* from $\mathcal{D}_{pt}$.
    ii. Compute $S(\mathcal{D}_{pt}^{(b)})$.
b. The empirical distribution of $\{S(\mathcal{D}_{pt}^{(b)})\}_{b=1}^{B}$ approximates the sampling distribution of $S(\mathcal{D}_{pt})$.

We typically set $B = 1000$ as a default. We set a common benchmark size n^* across sources (default $n^* = 250$) so that synthetic platforms do not appear more "stable" solely because they have larger raw sample sizes.

C.3. Alignment Pass-Rate Benchmarking (Synthetic vs. Human)

To summarize how often a synthetic platform reproduces a human benchmark within a tolerance band, we use a *pass-rate* metric inspired by recent work evaluating the statistical realism of LLM-generated data (Xie et al. 2026).

Let s denote a synthetic platform and h a human reference platform. At wave t, define a tolerance τ for the statistic S (e.g., standardized mean difference, proportion difference, Fisher-z difference). For each bootstrap replication:

$$\text{Pass}_{s,h,t}^{(b)} = \mathbf{1}\left(\left|S(\mathcal{D}_{st}^{(b)}, \mathcal{D}_{ht}^{(b)})\right| \le \tau\right).$$

For example, for the standardized mean difference (Cohen's d), we have:

$$d^{(b)} = S(\mathcal{D}_{st}^{(b)}, \mathcal{D}_{ht}^{(b)}) = \frac{\text{mean}(\mathcal{D}_{st}^{(b)}) - \text{mean}(\mathcal{D}_{ht}^{(b)})}{\text{sd}_{\text{pooled}}^{(b)}}$$

For correlation, we first compute the Fisher transformation of correlation r before analyzing the differences: $z(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$.

Finally, the estimated pass rate is

$$\widehat{\pi}_{s,h,t} = \frac{1}{B}\sum_{b=1}^{B} \text{Pass}_{s,h,t}^{(b)}.$$

We interpret higher pass rates as stronger agreement between synthetic and human samples *at the level of the selected statistic*, while emphasizing that pass rates depend on (i) the chosen tolerance τ, (ii) the resample size n^*, and (iii) the statistic itself.

For comparisons of Cronbach's α values, we also use α "hit-alignment" as an alternative

to pass rate. We define an *alpha-hit* indicator using a conventional threshold $\alpha_0 = 0.80$. In each bootstrap draw b, define hit indicators

$$H_{st}^{(b)} = \mathbf{1}\left(\widehat{\alpha}_{st}^{(b)} \geq \alpha_0\right), \qquad H_{ht}^{(b)} = \mathbf{1}\left(\widehat{\alpha}_{ht}^{(b)} \geq \alpha_0\right).$$

The α-hit *alignment* indicator is

$$\text{HitAlign}_{\alpha}^{(b)} = \mathbf{1}\left(H_{st}^{(b)} = H_{ht}^{(b)}\right),$$

i.e., the two groups either *both* meet the α_0 threshold or *both* fall below it in draw b.

The corresponding α-hit alignment rate is

$$\widehat{\pi}_{\alpha,\text{hit}} = \frac{1}{B}\sum_{b=1}^{B} \text{HitAlign}_{\alpha}^{(b)}.$$

C.4. Consistency Pass-Rate Benchmarking (Across Waves)

We apply the same logic as before for within-platform comparisons over time. For example, a per-bootstrap comparison of wave $t = 1$ and $t' = 2$ for the human sample is:

$$\text{Pass}_{h,t,t'}^{(b)} = \mathbf{1}\left(\left|S(\mathcal{D}_{ht}^{(b)}, \mathcal{D}_{ht'}^{(b)})\right| \leq \tau\right).$$

C.5. Tolerances and Link to Detectability.

In the main analyses, we tie tolerances to MDE-style thresholds used in the pre-registration (`https://aspredicted.org/5683tw.pdf`), i.e., we treat a difference as "small relative to detectability" if it falls below the approximate minimum detectable effect size (80% power, $\alpha = 0.05$) for the relevant n^*. This provides a transparent and scale-consistent way to interpret whether apparent drift is large enough to be practically consequential.

- *Standardized mean difference (Cohen's* d*):* For the MDE in standardized units (two-sample test), we use the following approximation:

$$\tau_d \approx \left(z_{1-\alpha/2} + z_{1-\beta}\right)\sqrt{\frac{2}{n}}.$$

  For $n^* = 250$ we have $\tau_d \approx (1.96 + 0.84) \times \sqrt{\frac{2}{250}} = 0.25$.
- *Proportion difference:* For two independent proportions with equal sample sizes, the

standard error of Δp is:

$$SE(\Delta p) = \sqrt{\frac{p_A(1-p_A)}{n} + \frac{p_B(1-p_B)}{n}}.$$

For unknown p, the conservative ("worst case") choice is $p_A \approx p_B \approx 0.5$, giving:

$$SE_{\max}(\Delta p) \approx \sqrt{\frac{0.25}{n} + \frac{0.25}{n}} = \sqrt{\frac{0.5}{n}}.$$

For $n^* = 250$ we have $\tau_p \approx (1.96 + 0.84) \times \sqrt{0.5/250} = 0.125$.

- *Fisher-z difference for (tetrachoric) correlations:* For independent samples, $z(r)$ is approximately normal with:

$$SE(z(r)) \approx \frac{1}{\sqrt{n-3}}.$$

So the SE for a difference is:

$$SE(\Delta z) \approx \sqrt{\frac{1}{n-3} + \frac{1}{n-3}} = \sqrt{\frac{2}{n-3}}$$

For $n^* = 250$ we have $\tau_z \approx (1.96 + 0.84) \times \sqrt{\frac{2}{250-3}} = 0.25$.
Tetrachoric correlations are often noisier, and here we rely on the bootstrap approximation to the standard error,

$$\widehat{SE}(\Delta z) = \mathrm{sd}\left(\Delta_z^{(1)}, \ldots, \Delta_z^{(B)}\right).$$

- *Average marginal effects (AMEs):* For regression-based estimands (e.g., average marginal effects in a logit/probit/linear model), we define $S(\cdot)$ as the AME computed after fitting the model on the bootstrap sample. Let $\widehat{\mathrm{AME}}_{st}^{(b)}$ and $\widehat{\mathrm{AME}}_{ht}^{(b)}$ denote the AME in bootstrap draw b for the synthetic and human samples, respectively. We compare AMEs via the bootstrap distribution of differences

$$\Delta_{\mathrm{AME}}^{(b)} = \widehat{\mathrm{AME}}_{st}^{(b)} - \widehat{\mathrm{AME}}_{ht}^{(b)}.$$

We set the tolerance τ_{AME} using the bootstrap-based standard error of the difference. Specifically, let

$$\widehat{SE}(\Delta AME) = \mathrm{sd}\left(\Delta_{\mathrm{AME}}^{(1)}, \ldots, \Delta_{\mathrm{AME}}^{(B)}\right).$$

We then define an MDE-style tolerance (80% power, $\alpha = 0.05$) as

$$\tau_{\text{AME}} = (1.96 + 0.84) \times \widehat{SE}(\Delta AME).$$

A bootstrap draw b is counted as a pass if $|\Delta^{(b)}_{\text{AME}}| \leq \tau_{\text{AME}}$.

- α *difference:* For each bootstrap draw b, compute Cronbach's α in the synthetic and human resamples, denoted $\widehat{\alpha}^{(b)}_{st}$ and $\widehat{\alpha}^{(b)}_{ht}$. Define the per-draw difference

$$\Delta^{(b)}_{\alpha} = \widehat{\alpha}^{(b)}_{st} - \widehat{\alpha}^{(b)}_{ht}.$$

Using the bootstrap approximation to the standard error,

$$\widehat{SE}(\Delta\alpha) = \text{sd}\left(\Delta^{(1)}_{\alpha}, \ldots, \Delta^{(B)}_{\alpha}\right),$$

we define an MDE-style tolerance (80% power, $\alpha = 0.05$) as

$$\tau_{\alpha} = (1.96 + 0.84) \times \widehat{SE}(\Delta\alpha).$$

A bootstrap draw b is counted as a pass if $|\Delta^{(b)}_{\alpha}| \leq \tau_{\alpha}$.

D. Survey

Section 1: Demographics

Before we start, please answer a few general questions about your personal information to help us better understand our respondents.

Q1. How would you describe your gender?

- Male
- Female
- Non-binary / third gender
- Prefer not to say

Q2. How old are you? (in years) _____

Q3. What is your ethnicity? (simplified)

- Asian
- Black
- Mixed
- White
- Other

Q4. What is the highest level of education you have completed?

- Some high school or less
- High school diploma or GED
- Some college, but no degree
- Associate's or technical degree
- Bachelor's degree
- Graduate or professional degree (MA, MS, PhD, JD, MD, DDS, etc.)
- Prefer not to say

Section 2: Consideration Factors

Subsequently, we would like to learn more about your personal thoughts and experiences with sustainable foods. Please take your time to think about each question and, if required, answer in your own words. There are no right or wrong answers, and we value your honest opinion.

Q5. For non-perishable foods, please write down the factors that you take into consideration when choosing a product within this product category. Please list from 2 to 6 unique factors, separating each entry with a comma.

Q6. For perishable foods, please write down the factors that you take into consideration when choosing a product within this product category. Please list from 2 to 6 unique factors, separating each entry with a comma.

Q7. For personal care products, please write down the factors that you take into consideration when choosing a product within this product category. Please list from 2 to 6 unique factors, separating each entry with a comma.

Section 3: Product Expectations

For each of the following statements, please indicate your level of agreement by selecting the number that best represents your opinion on a scale from 1 to 7.

Q8. How tasty would you rate plant-based meat alternatives? (not at all tasty / extremely tasty)
Q9. How healthy would you rate plant-based meat alternatives? (not at all healthy / extremely healthy)
Q10. How sustainable would you rate plant-based meat alternatives? (not at all sustainable / extremely sustainable)
Q11. How filling would you rate plant-based meat alternatives? (not at all filling / extremely filling)
Q12. How natural would you rate the ingredients of plant-based meat alternatives? (not at all natural / extremely natural)
Q13. How would you rate the price-to-quality ratio for plant-based meat alternatives? (extremely poor value / extremely good value)
Q14. How confident would you feel about the ingredients of plant-based meat alternatives? (highly suspicious / highly reassured)

Section 4: Product Engagement

For each of the following statements, please indicate your level of agreement by selecting the number that best represents your opinion on a scale from 1 to 7 (1 = extremely unlikely; 7 = extremely likely).

Q15. How likely would you be to . . .

. . . recommend plant-based meat alternatives to a friend?

. . . give plant-based meat alternatives a try?

. . . order plant-based meat alternatives?

. . . buy plant-based meat alternatives?

. . . choose plant-based meat alternatives over animal-based alternatives?

. . . pay a little more for plant-based meat alternatives?

Q16. In the past 30 days, how often did you eat plant-based meat alternatives?

- 0 times
- 1 time
- 2–3 times
- 4–6 times
- 7 times or more

Section 5: Burger Preference

Q17. We next present you with a selection of 5 distinct burger types. Each burger includes a standard set of ingredients: lettuce, tomato, mayo, burger sauce, and a brioche bun. They differ only by the **type of patty** used. Which of these burgers below do you prefer the most? *(single-select, randomized order)*

- **Beef Burger** (animal protein)
- **Tastes-Like-Meat Burger** (plant protein)
- **Veggie Burger** (plant protein)
- **Falafel Burger** (plant protein)
- **Chicken Burger** (animal protein)

Section 6: Burger Consideration Set

Q18. For the same set of burgers, please select any and all burgers that you could imagine ordering at a restaurant. *(multi-select, randomized order)*

- **Beef Burger** (animal protein)
- **Tastes-Like-Meat Burger** (plant protein)
- **Veggie Burger** (plant protein)
- **Falafel Burger** (plant protein)
- **Chicken Burger** (animal protein)

Section 7: Qualitative Items

Please take your time to think about each question and answer in your own words. There are no right or wrong answers, and we value your honest opinions.

Q19. What are, in your opinion, the top 3 unique reasons for buying plant-based meat alternatives?

Reason 1: _____

Reason 2: _____

Reason 3: _____

Q20. What are, in your opinion, the top 3 unique reasons against buying plant-based meat alternatives?

Reason 1: _____

Reason 2: _____

Reason 3: _____

Q21. The last time you tried a plant-based meat alternative, how would you describe the eating experience?

Q22. What would you like to see improved about plant-based meat alternatives?

Q23. Recall the last time you ate plant-based meat alternatives: please describe the situation.

Q24. How was your last shopping trip to the grocery store? Please describe your personal experience.

Section 8: Food Intuitions

For each of the following statements, please indicate your level of agreement by selecting the number that best represents your opinion on a scale from 1 to 7. (1 = completely disagree; 7 = completely agree).

Q25. Food that is *sustainable* is generally . . .

(a) . . . healthier.

(b) . . . tastier.

(c) . . . more expensive.

(d) . . . less filling.

(e) . . . more masculine.

(f) . . . more feminine.

Q26. Food that *tastes better* is generally . . .

(a) . . . healthier.

(b) . . . more sustainable.

(c) . . . more expensive.

(d) . . . less filling.

(e) . . . more masculine.

(f) . . . more feminine.

Section 9: Ambivalent Sexism Scale

The following section includes a variety of statements reflecting different opinions currently held in society regarding social roles and gender relations. Please indicate how much you personally agree or disagree with each statement. There are no right or wrong answers; we are simply interested in your honest perspective. (1 = completely disagree; 7 = completely agree).

Q27.

(a) Many women are actually seeking special favors, such as hiring policies that favor them over men, under the guise of asking for "equality."
(b) Most women interpret innocent remarks or acts as being sexist.
(c) Women are too easily offended.
(d) Feminists are NOT seeking for women to have more power than men.
(e) Most women fail to appreciate fully all that men do for them.
(f) Women seek to gain power by getting control over men.
(g) Women exaggerate problems they have at work.
(h) Once a woman gets a man to commit to her, she usually tries to put him on a tight leash.
(i) When women lose to men in a fair competition, they typically complain about being discriminated against.

(j) There are actually very few women who get a kick out of teasing men by seeming sexually available and then refusing male advances.
(k) Feminists are making entirely reasonable demands of men.

Section 10: Eating Habits

Q28. Which of the following best describes your diet?

- Omnivore (I regularly eat meat/dairy products)
- Flexitarian (I occasionally eat meat/dairy products)
- Pescetarian (I never eat meat, but I eat fish)
- Vegetarian (I never eat meat, but I eat dairy products)
- Vegan (I never eat animal-derived products)
- Halal/Kosher (I eat meat prepared according to Muslim/Jewish regulations)

Q29. Are you currently following a diet or are actively engaged in weight management?

- Yes
- No
- Prefer not to say

Q30. Before today, how familiar were you with plant-based meat alternatives? *(not at all familiar / completely familiar)*

Q31. How frequently do you eat meat?

- Never
- Rarely
- One to three times a month
- One to four times a week
- Every day or almost every day
- Multiple times a day

Q32. How frequently do you eat plant-based meat alternatives?

- Never
- Rarely
- One to three times a month
- One to four times a week
- Every day or almost every day
- Multiple times a day

Q33. Which of the following statements best reflects your eating behavior regarding meat?

- I am not interested in lowering my meat intake and I have no intention of doing that within the next six months.
- I am interested in lowering my meat intake and I have the intention of doing that within the next six months.
- I am interested in lowering my meat intake and I have the intention of doing that within the next month.
- I am interested in lowering my meat intake and I have started lowering my meat intake during the last six months.
- I am interested in lowering my meat intake and I have already lowered my meat intake for longer than six months.

Section 11: Fluid Intelligence (Numeracy); only for human sample

Please answer the questions below. Do not use a calculator, but feel free to take notes.

Q34. Imagine that we flip a fair coin 1,000 times. What is your best guess about how many times the coin would come up heads in 1,000 flips?

Q35. In the BIG BUCKS LOTTERY, the chance of winning a $10 prize is 1%. What is your best guess about how many people would win a $10 prize if 1,000 people each buy a single ticket to BIG BUCKS?

Q36. In ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1,000. What percent of tickets to ACME PUBLISHING SWEEPSTAKES win a car?

Section 12: Prior Experience; only for human sample

Q37. Have you previously encountered the specific math and logic problems you just completed on the previous page ("numeracy test")?
- Yes
- No
- Not sure

Q38. Before today, how often have you taken surveys about plant-based meat alternatives?
- Never
- Once
- A few times
- Many times

Q39. Have you previously answered questions about whether sustainable food is “more masculine/feminine”?

- Yes
- No
- Not sure

E. Variables

TABLE 1. Overview of Benchmarking Variables

Nr.	Variable	Category	Type	Subset
1	Age	Demographics	Numerical	Yes
2	Analog burger consideration	Preferences	Proportion	Yes
3	Analog burger preference	Preferences	Proportion	Yes
4	Analog-chicken burger consideration	Relational	Tetrachoric Correlation	No
5	Analog-non-analog burger consideration	Relational	Tetrachoric Correlation	No
6	Analog-semi-analog burger consideration	Relational	Tetrachoric Correlation	No
7	Animal-based burger consideration	Preferences	Proportion	No
8	Animal-based burger preference	Preferences	Proportion	No
9	Animal-based diet	Consumer Characteristics	Proportion	Yes
10	Beef burger consideration	Preferences	Proportion	Yes
11	Beef burger preference	Preferences	Proportion	Yes
12	Beef-analog burger consideration	Relational	Tetrachoric Correlation	Yes
13	Beef-chicken burger consideration	Relational	Tetrachoric Correlation	No
14	Beef-non-analog burger consideration	Relational	Tetrachoric Correlation	No
15	Beef-semi-analog burger consideration	Relational	Tetrachoric Correlation	No
16	Behavior change stage (interested)	Consumer Characteristics	Proportion	Yes
17	Burger preference entropy	Preferences	Numerical	Yes
18	Chicken burger consideration	Preferences	Proportion	No
19	Chicken burger preference	Preferences	Proportion	No
20	Chicken-non-analog burger consideration	Relational	Tetrachoric Correlation	No
21	Chicken-semi-analog burger consideration	Relational	Tetrachoric Correlation	No
22	Education (Bachelor or higher)	Demographics	Proportion	Yes
23	Engagement	Reliability	Alpha	Yes
24	Engagement	Consumer Characteristics	Numerical	Yes
25	Engagement ~ Tasty (expectations)	Relational	AME	Yes
26	Engagement ~ Healthy (expectations)	Relational	AME	No
27	Engagement ~ Sustainable (expectations)	Relational	AME	Yes
28	Engagement ~ Filling (expectations)	Relational	AME	No
29	Engagement ~ Naturalness (expectations)	Relational	AME	No
30	Engagement ~ Price (expectations)	Relational	AME	Yes
31	Engagement ~ Ingredients (expectations)	Relational	AME	No
32	Engagement ~ Healthy (intuitions)	Relational	AME	No
33	Engagement ~ Tasty (intuitions)	Relational	AME	No
34	Engagement ~ Price (intuitions)	Relational	AME	No
35	Engagement ~ Filling (intuitions)	Relational	AME	No
36	Engagement ~ Masculinity (intuitions)	Relational	AME	No
37	Engagement ~ Femininity (intuitions)	Relational	AME	No
38	Ethnicity (White)	Demographics	Proportion	Yes
39	Gender (female)	Demographics	Proportion	Yes
40	Heavy meat eater	Consumer Characteristics	Proportion	Yes
41	Heavy PBMA eater	Consumer Characteristics	Proportion	No
42	Never meat eater	Consumer Characteristics	Proportion	No

Continued on next page

TABLE 1 – continued from previous page

Nr.	Variable	Category	Type	Subset
43	Never PBMA eater	Consumer Characteristics	Proportion	No
44	Non-analog burger consideration	Preferences	Proportion	Yes
45	Non-analog burger preference	Preferences	Proportion	Yes
46	PBMA familiarity	Consumer Characteristics	Numerical	No
47	PBMA healthiness	Product Perceptions	Numerical	No
48	PBMA ingredients	Product Perceptions	Numerical	No
49	PBMA naturalness	Product Perceptions	Numerical	No
50	PBMA price-to-quality	Product Perceptions	Numerical	No
51	PBMA purchases last month	Consumer Characteristics	Numerical	No
52	PBMA satiety	Product Perceptions	Numerical	No
53	PBMA sustainability	Product Perceptions	Numerical	Yes
54	PBMA tastiness	Product Perceptions	Numerical	Yes
55	Plant-animal burger consideration	Relational	Tetrachoric Correlation	Yes
56	Plant-based burger consideration	Preferences	Proportion	No
57	Plant-based diet	Consumer Characteristics	Proportion	No
58	$Pr(\text{Animal} = 1) \sim$ Sexism	Relational	AME	No
59	$Pr(\text{Animal} = 1) \sim$ Filling	Relational	AME	No
60	$Pr(\text{Animal} = 1) \sim$ Healthy	Relational	AME	No
61	$Pr(\text{Animal} = 1) \sim$ Ingredients	Relational	AME	No
62	$Pr(\text{Animal} = 1) \sim$ Naturalness	Relational	AME	No
63	$Pr(\text{Animal} = 1) \sim$ Price	Relational	AME	Yes
64	$Pr(\text{Animal} = 1) \sim$ Sustainable	Relational	AME	Yes
65	$Pr(\text{Animal} = 1) \sim$ Tasty	Relational	AME	Yes
66	Readability (eating experience)	Product Perceptions	Numerical	Yes
67	Readability (improvements)	Consumer Characteristics	Numerical	No
68	Readability (shopping trip)	Consumer Characteristics	Numerical	No
69	Readability (situation)	Product Perceptions	Numerical	No
70	Semi-analog burger consideration	Preferences	Proportion	No
71	Semi-analog burger preference	Preferences	Proportion	No
72	Semi-analog–non-analog burger consideration	Relational	Tetrachoric Correlation	No
73	Sexism	Reliability	Alpha	Yes
74	Sexism	Consumer Characteristics	Numerical	Yes
75	Sustainability expensive–healthy	Relational	Correlation	No
76	Sustainability expensive–tasty	Relational	Correlation	No
77	Sustainability feminine–expensive	Relational	Correlation	No
78	Sustainability feminine–healthy	Relational	Correlation	No
79	Sustainability feminine–masculine	Relational	Correlation	No
80	Sustainability feminine–tasty	Relational	Correlation	No
81	Sustainability filling–expensive	Relational	Correlation	No
82	Sustainability filling–feminine	Relational	Correlation	No
83	Sustainability filling–healthy	Relational	Correlation	No
84	Sustainability filling–masculine	Relational	Correlation	No
85	Sustainability filling–tasty	Relational	Correlation	No
86	Sustainability masculine–expensive	Relational	Correlation	No
87	Sustainability masculine–healthy	Relational	Correlation	No
88	Sustainability masculine–tasty	Relational	Correlation	No
89	Sustainability tasty–healthy	Relational	Correlation	No

Continued on next page

TABLE 1 – continued from previous page

Nr.	Variable	Category	Type	Subset
90	Sustainability=Expensive	Product Perceptions	Numerical	No
91	Sustainability=Feminine	Product Perceptions	Numerical	Yes
92	Sustainability=Filling	Product Perceptions	Numerical	No
93	Sustainability=Healthy	Product Perceptions	Numerical	No
94	Sustainability=Masculine	Product Perceptions	Numerical	No
95	Sustainability=Tasty	Product Perceptions	Numerical	No
96	Taste expensive–healthy	Relational	Correlation	No
97	Taste expensive–sustainable	Relational	Correlation	No
98	Taste feminine–expensive	Relational	Correlation	No
99	Taste feminine–filling	Relational	Correlation	No
100	Taste feminine–healthy	Relational	Correlation	No
101	Taste feminine–masculine	Relational	Correlation	No
102	Taste feminine–sustainable	Relational	Correlation	No
103	Taste filling–expensive	Relational	Correlation	No
104	Taste filling–healthy	Relational	Correlation	No
105	Taste filling–sustainable	Relational	Correlation	No
106	Taste masculine–expensive	Relational	Correlation	No
107	Taste masculine–filling	Relational	Correlation	No
108	Taste masculine–healthy	Relational	Correlation	No
109	Taste masculine–sustainable	Relational	Correlation	No
110	Taste sustainable–healthy	Relational	Correlation	No
111	Taste=Sustainable–Sustainability=Tasty	Relational	Correlation	No
112	Tasty=Expensive	Product Perceptions	Numerical	No
113	Tasty=Feminine	Product Perceptions	Numerical	No
114	Tasty=Filling	Product Perceptions	Numerical	No
115	Tasty=Healthy	Product Perceptions	Numerical	Yes
116	Tasty=Masculine	Product Perceptions	Numerical	No
117	Tasty=Sustainable	Product Perceptions	Numerical	No
118	Text Consistency	Relational	Numerical	No
119	Weight management	Consumer Characteristics	Proportion	Yes
120	Words used (eating experience)	Product Perceptions	Numerical	No
121	Words used (improvements)	Consumer Characteristics	Numerical	No
122	Words used (shopping trip)	Consumer Characteristics	Numerical	No
123	Words used (situation)	Product Perceptions	Numerical	No

F. Distribution of Demographics across Samples

TABLE 2. Distributions across Demographic Cells by Wave.

Panel A: Wave 1

		Single-Prompt Batch				Iterative Generation			Seed-And-Expand
Sample	U.S. %	Prolific %	ChatGPT %	Qwen %	Gemini %	ChatGPT %	Qwen %	Gemini %	Gemini %
18–24, f, low edu	4.79	4.07	8.50	3.28	6.79	6.83	2.54	5.19	5.51
18–24, f, high edu	0.93	0.37	0.00	0.23	0.00	0.00	0.00	0.00	1.27
18–24, m, low edu	5.31	2.22	4.75	7.49	7.07	6.83	1.95	4.36	6.57
18–24, m, high edu	0.68	0.00	0.50	0.23	0.00	0.00	0.00	0.00	1.06
25–44, f, low edu	9.27	28.89	18.25	15.93	20.11	16.61	15.23	19.50	21.40
25–44, f, high edu	7.81	11.11	3.00	1.41	3.80	4.98	11.33	3.53	3.81
25–44, m, low edu	10.99	17.78	13.25	18.50	11.68	11.99	10.94	18.88	18.64
25–44, m, high edu	6.43	2.59	3.25	6.09	7.34	0.55	1.95	2.70	4.87
45–64, f, low edu	9.76	11.48	9.50	11.71	10.87	24.72	14.26	14.32	11.65
45–64, f, high edu	5.88	2.22	3.75	5.15	0.27	4.43	24.41	2.49	3.18
45–64, m, low edu	10.09	8.89	13.75	8.67	11.14	10.52	10.55	13.69	13.14
45–64, m, high edu	5.16	3.33	2.75	4.68	7.88	0.92	2.54	3.11	3.81
65+, f, low edu	8.92	4.07	9.50	9.60	4.35	8.30	1.56	5.39	2.12
65+, f, high edu	3.64	0.74	0.25	0.94	3.26	0.18	0.39	1.24	0.85
65+, m, low edu	6.74	1.85	7.25	5.62	4.89	3.14	1.76	4.36	1.69
65+, m, high edu	3.62	0.37	1.75	0.47	0.54	0.00	0.59	1.24	0.42

Panel B: Wave 2

		Single-Prompt Batch				Iterative Generation			Seed-And-Expand
Sample	U.S. %	Prolific %	ChatGPT %	Qwen %	Gemini %	ChatGPT %	Qwen %	Gemini %	Gemini %
18–24, f, low edu	4.79	5.03	6.72	1.73	7.19	6.85	2.92	5.32	4.21
18–24, f, high edu	0.93	0.00	0.00	0.00	0.00	0.00	0.18	0.00	1.33
18–24, m, low edu	5.31	4.03	3.95	4.32	15.55	4.81	2.37	4.29	4.66
18–24, m, high edu	0.68	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.44
25–44, f, low edu	9.27	24.83	14.43	10.37	20.42	16.30	27.74	19.63	18.63
25–44, f, high edu	7.81	8.39	6.13	9.07	5.10	4.81	14.05	2.66	4.88
25–44, m, low edu	10.99	20.47	16.01	16.20	19.72	15.19	14.42	17.79	19.51
25–44, m, high edu	6.43	3.36	2.96	1.08	1.16	1.11	2.55	3.27	5.99
45–64, f, low edu	9.76	16.11	11.46	12.10	6.96	17.04	14.60	15.13	14.41
45–64, f, high edu	5.88	2.68	3.95	7.56	6.50	5.19	6.57	2.66	4.43
45–64, m, low edu	10.09	7.38	15.42	15.33	8.58	12.41	7.85	13.91	13.97
45–64, m, high edu	5.16	1.34	2.17	1.30	3.02	2.78	2.92	1.64	2.00
65+, f, low edu	8.92	3.69	7.51	9.29	1.39	5.37	1.82	6.95	1.11
65+, f, high edu	3.64	0.67	0.79	0.22	1.62	0.74	0.73	0.41	0.44
65+, m, low edu	6.74	1.68	7.11	11.23	0.93	7.22	0.91	5.11	1.11
65+, m, high edu	3.62	0.34	1.38	0.22	1.86	0.19	0.36	1.23	0.89

Panel C: Wave 3

		Single-Prompt Batch				Iterative Generation			Seed-And-Expand
Sample	U.S. %	Prolific %	ChatGPT %	Qwen %	Gemini %	ChatGPT %	Qwen %	Gemini %	Gemini %
18–24, f, low edu	4.79	9.47	7.60	8.14	6.45	3.33	3.84	2.00	6.29
18–24, f, high edu	0.93	0.70	0.23	0.43	0.00	0.00	0.00	0.00	1.40
18–24, m, low edu	5.31	2.81	5.76	2.78	8.99	3.70	7.68	3.00	6.29
18–24, m, high edu	0.68	0.35	0.23	0.00	0.00	0.00	0.00	0.00	1.63
25–44, f, low edu	9.27	22.81	15.67	17.56	17.97	16.64	9.51	12.60	16.32
25–44, f, high edu	7.81	6.67	3.69	2.57	4.38	7.76	7.86	2.80	6.06
25–44, m, low edu	10.99	22.46	11.06	14.56	18.89	15.71	9.69	22.00	17.95
25–44, m, high edu	6.43	3.16	2.07	0.86	2.76	1.48	2.93	1.20	4.90
45–64, f, low edu	9.76	14.39	14.06	7.28	11.52	22.92	32.36	14.80	13.52
45–64, f, high edu	5.88	5.61	1.38	4.28	3.46	4.81	4.75	2.00	2.80
45–64, m, low edu	10.09	6.67	12.21	15.42	12.21	12.75	10.97	21.20	16.08
45–64, m, high edu	5.16	1.05	2.07	4.50	4.84	1.48	3.84	1.60	2.33
65+, f, low edu	8.92	2.46	11.52	7.71	2.76	8.13	3.11	5.00	1.40
65+, f, high edu	3.64	0.70	1.38	1.28	2.30	0.00	1.10	0.80	0.47
65+, m, low edu	6.74	0.70	8.99	9.85	0.69	1.29	1.28	9.00	2.33
65+, m, high edu	3.62	0.00	2.07	2.78	2.76	0.00	1.10	2.00	0.23

G. Results: Unweighted Data

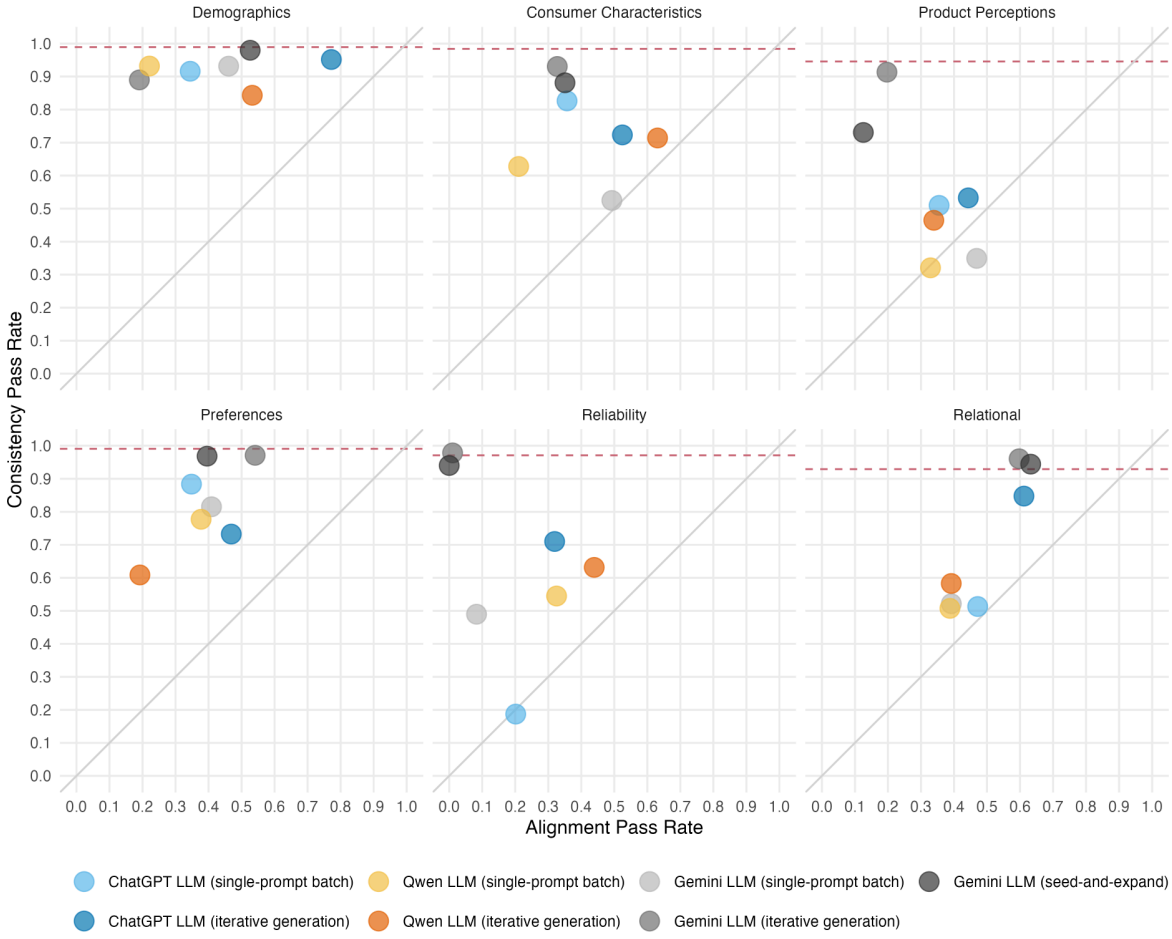

FIGURE 1. Comparison of Alignment and Consistency Pass Rates (Dashed Line = Prolific)

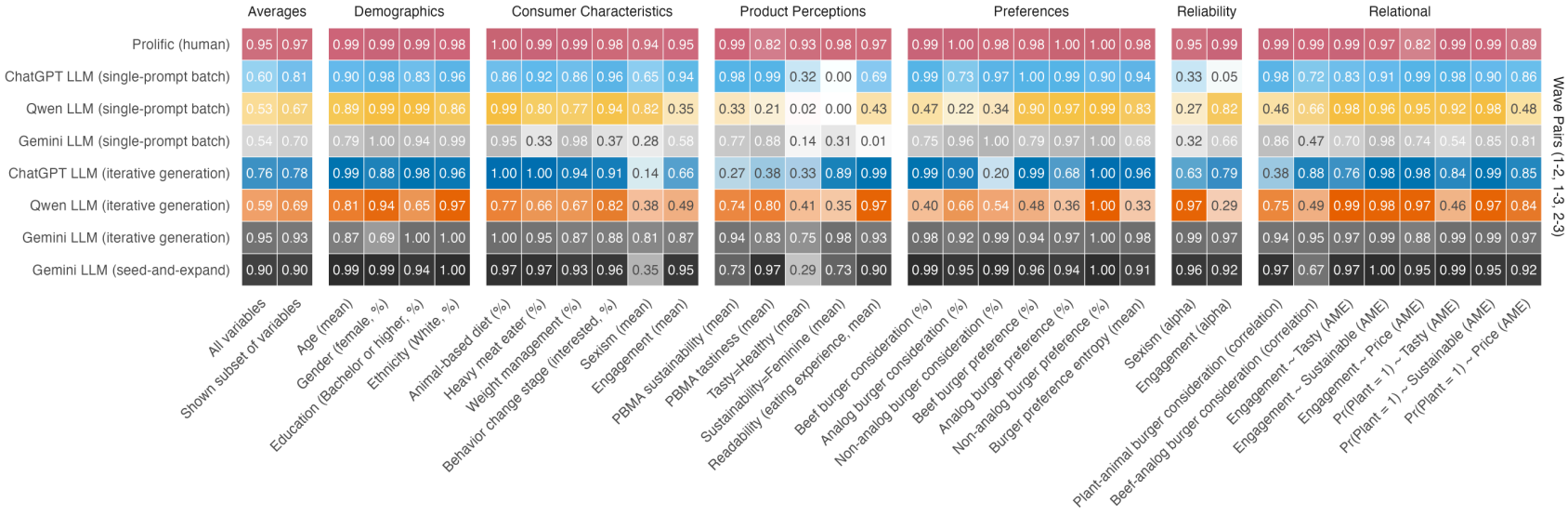

FIGURE 2. Consistency Pass Rate (Average across Wave Pairs within each Platform)

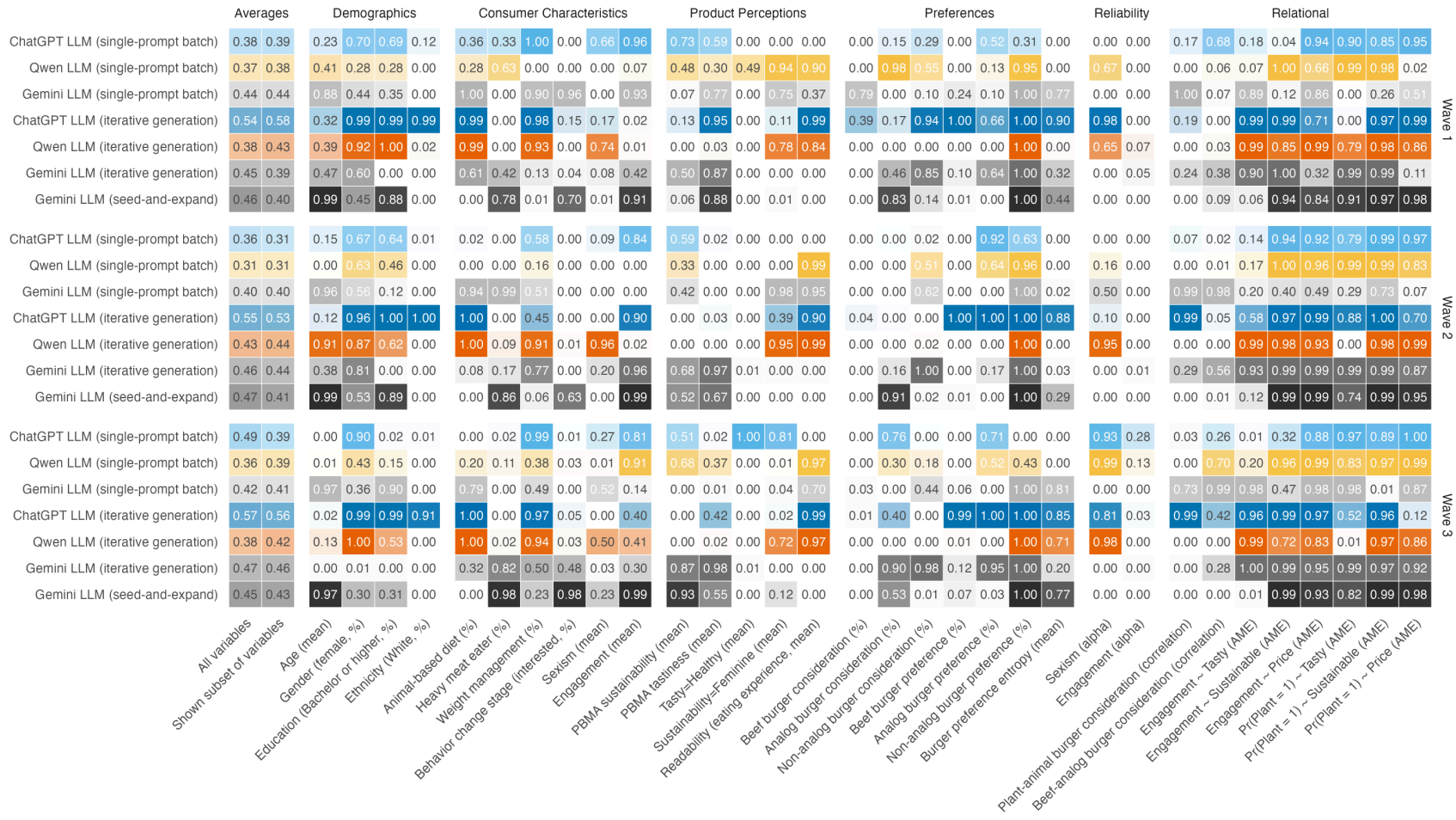

FIGURE 3. Alignment Pass Rates (Prolific as Benchmark in each Wave)

References

Davison, Anthony C., & David V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Efron, Bradley, & Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC.

Hall, Peter. 1992. *The Bootstrap and Edgeworth Expansion*. New York: Springer.

Xie, Yueqi, Lemeng Liang, Shuzhen Li, Yifu Lu, Zhiwen Xiao, Mengdi Shi, Junming Huang, Mengdi Wang, & Yu Xie. 2026. "Evaluating the statistical realism of LLM-generated social science data." *Proceedings of the National Academy of Sciences* 123 (19): e2538145123.