LLMs Reproduce Human Purchase Intent via Semantic Similarity Elicitation of Likert Ratings

Benjamin F. Maier*
ben.maier@pymc-labs.com
PyMC Labs
Tallinn, Estonia

Ulf Aslak
ulf.aslak@pymc-labs.com
PyMC Labs
Tallinn, Estonia

Luca Fiaschi
luca.fiaschi@pymc-labs.com
PyMC Labs
Tallinn, Estonia

Nina Rismal
nina.rismal@pymc-labs.com
PyMC Labs
Tallinn, Estonia

Kemble Fletcher
kemble.fletcher@pymc-labs.com
PyMC Labs
Tallinn, Estonia

Christian C. Luhmann
christian.luhmann@pymc-labs.com
PyMC Labs
Tallinn, Estonia

Robbie Dow
robbie_dow@colpal.com
Colgate-Palmolive Company
New York, NY, USA

Kli Pappas*
kli_pappas@colpal.com
Colgate-Palmolive Company
New York, NY, USA

Thomas V. Wiecki
thomas.wiecki@pymc-labs.com
PyMC Labs
Tallinn, Estonia

Abstract

Consumer research costs companies billions annually yet suffers from panel biases and limited scale. Large language models (LLMs) offer an alternative by simulating synthetic consumers, but produce unrealistic response distributions when asked directly for numerical ratings. We present semantic similarity rating (SSR), a method that elicits textual responses from LLMs and maps these to Likert distributions using embedding similarity to reference statements. Testing on an extensive dataset comprising 57 personal care product surveys conducted by a leading corporation in that market (9,300 human responses), SSR achieves 90% of human test–retest reliability while maintaining realistic response distributions (KS similarity > 0.85). Additionally, these synthetic respondents provide rich qualitative feedback explaining their ratings. This framework enables scalable consumer research simulations while preserving traditional survey metrics and interpretability.

CCS Concepts

• **Applied computing** → **Marketing**; • **Computing methodologies** → **Natural language generation**.

Keywords

Purchase Intent, Synthetic Consumers, Synthetic Focus Groups, NLP, LLMs

1 Introduction

Established consumer research plays a central role in guiding corporations' product development decisions [1–3], costing them billions of U.S. dollars globally every year [3]. Before investing in costly production and launch activities, companies routinely evaluate product concepts by surveying representative consumer panels. The most consequential question in such studies typically concerns purchase intent (PI), i.e., the likelihood that a respondent would buy the product if available [4–6]. Standard practice is to elicit purchase intent on a Likert scale, usually ranging from 1 (e.g., "definitely not") to 5 (e.g., "definitely yes") [7]. While widely used, this method faces well-known limitations: responses may be distorted by satisficing, acquiescence, and positivity biases, among other factors [8, 9]. Thus, traditional consumer panels often provide noisy measurements of demand, despite the considerable resources invested.

Recent advances in LLMs raise the possibility of augmenting or partially replacing human survey panels with synthetic consumers. By conditioning LLMs on demographic or attitudinal personas and exposing them to the same survey instruments, researchers have begun exploring whether such synthetic samples can recover human-like patterns of response. This line of work has expanded rapidly across disciplines, including market research, political science, psychology, and consumer behavior [10–15]. Taken together, this literature establishes the prominence of LLM-based synthetic samples while also underscoring challenges regarding their reliability.

One recurring challenge is the direct elicitation of Likert-scale responses. When asked to provide numerical ratings, LLMs tend to produce distributions that are overly narrow, systematically skewed, or otherwise inconsistent with human survey data [13–15]. This raises the question of whether such shortcomings reflect fundamental limits of LLMs as survey respondents, or simply the elicitation methods used.

In this paper, we argue for the latter. We show that the problem lies not in the use of LLMs themselves, but in directly requesting Likert-scale outputs. Instead, we propose to use textual elicitation followed by semantic-similarity rating (SSR): LLMs generate free-text statements of purchase intent, which are then projected onto a 5-point (5pt) Likert scale by computing the cosine similarity of embeddings with those of predefined anchor statements. This approach draws on established methods in NLP (semantic similarity mapping) [16] and survey methodology (anchoring vignettes) [17], but to our knowledge has not been applied in the context of LLMs as survey respondents.

Using 57 consumer research surveys on personal care product concepts conducted by a leading corporation in that market, each with 150–400 human participants, we demonstrate that SSR closely

*Corresponding authors.

replicates real survey outcomes. Specifically, it recovers both (1) the panel-level response distributions and (2) the relative ranking of product concepts by mean purchase intent. To quantify the latter, we introduce *correlation attainment* herein which is inspired by human test–retest reliability experiments and measures correlation between synthetic and real data in terms of its maximum achievable value. Moreover, we demonstrate that these results are only achieved when LLMs are prompted to consider demographic attributes of a person they are being asked to impersonate. We find that the LLMs' response behavior with regard to age and income level, in particular, mirrors actual humans' response behavior relatively well. As a byproduct of SSR, feedback on product concepts can be leveraged for qualitative analyses and further concept development.

These results suggest that, when paired with appropriate elicitation methods, LLMs can serve as valid synthetic consumers for concept testing.

2 Related Works

Several studies using LLMs as synthetic survey respondents rely on direct numeric elicitation. For example, models are asked to fill in Likert scales [15] or provide "feeling thermometer" scores [13]. Others adapt this approach to discrete-choice tasks, such as conjoint-style willingness-to-pay estimation [11] or behavioral games [18]. A consistent limitation is that response distributions are too narrow: models regress to "typical" answers, showing far less variance than human data and producing unrealistically confident estimates [13, 18].

A smaller line of work explores textual responses that are subsequently mapped onto numbers. One study uses elicitation of open completions ("[Brand] is similar to...") and converts those into similarity scores by counting elicited brand completions [10]. Another trains "Doppelgänger LLMs" on individual utterances, generating free-text survey answers that are then aligned with structured categories [19]. While such pipelines acknowledge the ambiguity of open-ended responses, they ultimately reduce them back to single numbers.

Another focus of some studies is demographic conditioning, where prompts embed socio-demographic backstories. One study shows that this improves the alignment of synthetic subgroup responses with human benchmarks [12], and another demonstrates similar effects with political personas [13]. Conditioning increases validity but does not overcome the fundamental issue of narrow distributions.

Some work uses fine-tuning with survey data to make LLMs more human-like [11, 12, 19]. But a large share of the literature, including refs. [10, 13, 18, 20], stays with zero-shot elicitation or prompt engineering.

3 Methods

This section provides a brief overview of the methods employed in this paper. See App. A for a detailed explanation.

3.1 Data

We analyze 57 consumer research surveys on personal care product concepts conducted by a leading corporation in that market using a digital consumer research platform (see App. A.1). Each survey involved 150–400 unique U.S. participants (N = 9,300 in total), with core demographic markers such as age, gender, and location available for most respondents, and income and ethnicity for fewer. Surveys asked participants to evaluate a concept and rate their purchase intent on a 5pt Likert scale. Mean purchase intent is skewed towards positive values and narrowly distributed with mean 4.0 and standard deviation 0.1 across all surveys.

3.2 Definitions

Following the definitions in App. A.2, each survey s is associated with a product concept and a set of consumers $c \in C_s$, who each provide a Likert rating $r_c \in \{1, \ldots, 5\}$ marking their purchase intent. Per survey, these form empirical response distributions and mean purchase intent PI_s. We define synthetic consumers c' as LLMs impersonating human respondents c given their demographic attributes. Unlike real consumers, their responses may be full probability distributions $p_{c'}(r)$. Throughout, we denote human data by superscript x and synthetic data by y.

3.3 Success Metrics

We evaluate synthetic panels using two main criteria (see App. A.3 for more detail):

Distributional Similarity. We measure per-survey similarity between synthetic and real purchase intent distributions via Kolmogorov–Smirnov (KS) similarity, i.e. $\mathrm{KS\ sim}_s^{xy} = 1 - \mathrm{KS\ dist}_s^{xy}$, because it respects the ordinality of the scale. For each experiment, we then report the mean KS similarity $K^{xy} = \mathrm{E}[\mathrm{KS\ sim}_s^{xy}]$ over all 57 surveys.

Relative Concept Appeal and Correlation Attainment. We compute Pearson correlations between mean purchase intents of real and synthetic surveys $R^{xy} = \mathrm{corr}[\mathrm{PI}^x, \mathrm{PI}^y]$ to quantify how well synthetic consumers recover relative concept appeal. Because correlation is upper bounded by noisy human data with a narrow PI_s distribution, we measure success across all 57 surveys in terms of the maximum attainable correlation, akin to test–retest reliability. For every experiment, we estimate this ceiling by simulating 2,000 test–retest scenarios, splitting each survey into two equally-sized cohorts for each scenario: test and control. Then, the maximum attainable correlation is given by R^{xx} between test and control cohorts. Correlation attainment is then quantified as $\rho = \mathrm{E}[R^{xy}]/\mathrm{E}[R^{xx}]$ where R^{xy} is the correlation between mean purchase intents of the test cohorts and equally-sized random samples of the corresponding synthetic surveys, respectively.

3.4 Synthetic Response Generation

Synthetic consumers were instantiated by prompting LLMs with demographic attributes and a product concept (App. A.4). If not stated otherwise we used the full concept image as a stimulus. We evaluated three response strategies (see Fig. 1A):

Direct Likert rating (DLR). LLMs reply with a single integer 1, 2, 3, 4, or 5.

Follow-up Likert rating (FLR). LLMs first generate a short free-text purchase intent statement, which is then mapped to a Likert score by a new instance of the same model which, this time, received a system prompt to act as a "Likert rating expert." In this

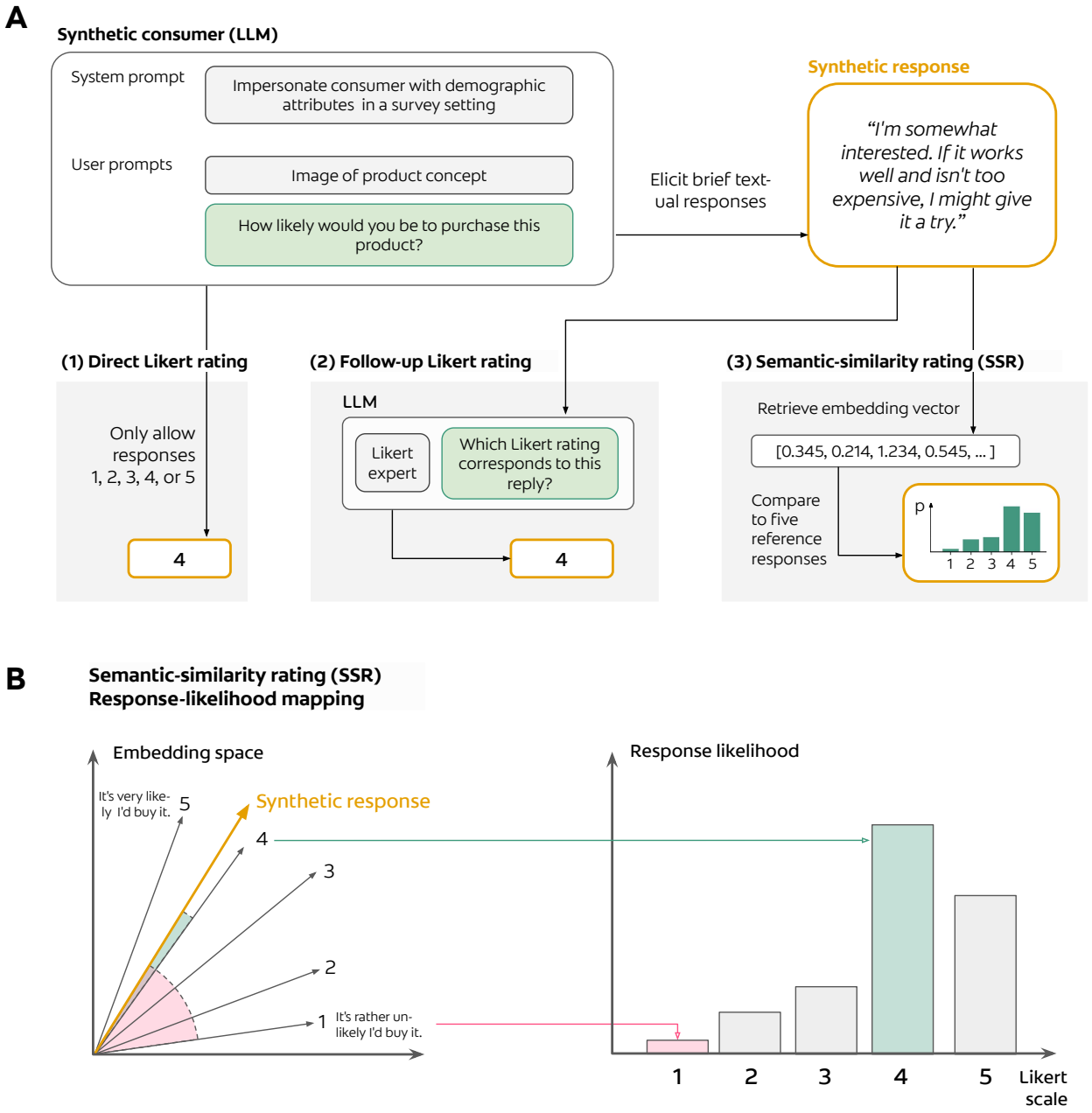

Figure 1: Different response generation procedures and SSR response-likelihood mapping. (A) A synthetic consumer is constructed by instructing an LLM to impersonate a consumer with certain demographic properties and show them a product concept as an image containing a description and possibly concept art (see App. B). The synthetic consumer is then asked about their purchase intent. (1) In the direct Likert-rating approach, the LLM's response is restricted to one of 1, 2, 3, 4, or 5. (2) Alternatively, we let the LLM write a brief textual response about their PI. Subsequently, we prompt the same model to be a Likert-rating "expert" and map the textual response to an integer between 1 and 5. (3) Because textual responses can result in varied ratings on the 5pt Likert scale, we introduce the semantic similarity rating method. We retrieve the embedding vector for the textual response from a corresponding model, compare it to five reference response embedding vectors and construct a response distribution on the Likert scale. (B) In an embedding space, the synthetic response will have a certain angular distance to any other statement. We construct a reference set of five rating responses, each corresponding to an integer on the Likert scale. Then, the response likelihood of any integer is set to be proportional to the cosine similarity between the synthetic response vector and the corresponding reference response vector.

system prompt, we included examples of what kind of statements can lead to which rating.

Semantic similarity rating (SSR). The same free-text responses are embedded and compared to reference statements for each point on the Likert scale, yielding a response probability mass function (pmf) with single probability values being proportional to the cosine similarity between the response and the corresponding reference statement (see Fig. 1B). Experiments reported in the main text used pmfs that were averaged over six different statement sets for every response (see App. C.1). Embedding vectors were retrieved using OpenAI's model "text-embedding-3-small."

We used two models (GPT-4o and Gemini-2.0-flash, "Gem-2f" in the following) and ran experiments with $T_{\text{LLM}} = 0.5$ and $T_{\text{LLM}} = 1.5$. As there was little variation between experiments at different temperatures, we only report results for $T_{\text{LLM}} = 0.5$ in the main text.

4 Results

4.1 Direct Likert Ratings

To establish a baseline for comparison, we first tested the performance of asking synthetic consumers for a Likert rating directly, using full information on demographic attributes. Both LLMs yielded a correlation attainment of about $\rho = 80\%$ (cf. Fig. 2A.i and Fig. 6A.i). At the same time, distributional similarity was poor with a mean similarity of $K^{xy} = 0.26$ for GPT-4o and $K^{xy} = 0.39$ for Gem-2f (cf. Figs. 2B, 3, 6B, 3, and 13–16). Upon detailed inspection of the Likert response distributions, models typically replied with response '3', i.e. a "safe" regression to the center of the scale (cf. Figs 9–12). The comparably high correlation with real data was therefore purely a result of occasional responses '2' and '4'. Almost never did the models reply with '1' or '5'. In contrast, the most responses in the real data were values '4' and '5'. Subsequent attempts to nudge models to explore the upper extremes of the distribution via system prompt modification lead to slightly higher distributional similarities while decreasing correlation in mean purchase intent, however. I.e. models then "over-corrected" in the direction of distributional similarity to real data, resulting in a loss of signal in product ranking, which subverts the overall goal of obtaining useful information about consumer purchase intent.

4.2 Textual Elicitation with FLR and SSR

Letting LLMs respond freely and using the generated responses to obtain Likert ratings yields correlation attainment values of $\rho = 85\%$ for GPT-4o (see Fig. 2A.ii) and $\rho = 90\%$ for Gem-2f (see Fig. 6A.ii). With GPT-4o consumers, FLRs achieve slightly lower correlation attainment than SSRs with $\rho = 90\%$ (see Fig. 2A.iii). With Gem-2f, both methods reach similar values (see Fig. 6A.iii). For the SSR method, distributional similarity markedly increased compared to the naive DLR approach, with $K^{xy} = 0.88$ for GPT-4o (see Fig. 3) and $K^{xy} = 0.8$ for Gem-2f (see Fig. 7). FLRs yield improved distributions compared to DLRs, but fall behind distributional similarity values reached by SSRs, ($K^{xy} = 0.72$ for GPT-4o and $K^{xy} = 0.59$ for Gem-2f, respectively, see Figs. 3 and 7). For this rating method, we found that equipping the system prompt with explicit examples of what kind of sentiments may lead to which rating on the Likert scale was necessary to avoid the narrow distributions observed with the DLR approach.

Generally, the synthetic mean purchase intents are far more spread out than the real mean purchase intents: When a product is less attractive, LLMs tend to rate them lower than their human counter parts, on average. For detailed results, see Figs. 17–22 and Tab. 1.

4.3 Influence of Demographics and Concept Properties

Furthermore, we are interested to find out which other aspects of the real survey data are mirrored by synthetic consumers (SCs). To this end, we measure mean purchase intent across all products, stratified by demographics and product features and present the results in Fig. 4.

Mean purchase intent follows a concave behavior with regards to age: both younger and older participants tended to rate their purchase intent lower than middle-aged age cohorts. This behavior is mirrored by synthetic consumers based on GPT-4o. For Gem-2f, younger synthetic consumers reported lower purchase intent. Older consumers, however, reported similar purchase intent as their middle-aged counter parts (see Fig. 4A).

In the real surveys, consumers had to rate their income level according to one of six reference statements. Here, statements 1 through 4 all suggested budgetary problems. Hence, it is unsurprising that only for statements 5 and "Null" (i.e. "None of these") there is a marked increase in purchase intent. This behavior is replicated both by GPT-4o and Gem-2f (see Fig. 4B): Personas prompted to have budgetary problems responded with lower purchase intent. GPT-4o reacted very sensitively to being prompted with income level 2, potentially due to the statement's drastic wording of being "in danger."

Both humans and SCs rated "Cat. IV" products consistently high and those from "Cat. I" consistently the lowest (see Fig. 4C). Moreover, humans and SCs alike reacted negatively to concepts developed by "Source B" (see Fig. 4D). Regarding a product's price segment, SCs replicated human behavior rather well once again, rating products from 'Tier 3' and 'Tier 4' more positively and 'Tier 1' lowest (see Fig. 4E).

SCs replicated the response behavior less well for gender and dwelling region (see Fig. 8). However, mean purchase intent is not being influenced strongly by those features for neither real nor any of the synthetic surveys.

To explore how well models leveraged the information contained in demographic attributes to arrive at the results above, we ran an additional experiment using Gem-2f and an SC system prompt that left out all demographic features. Surprisingly, this resulted in survey distributions that were very close to the real distributions, with consistently high purchase intent of '4' and '5' and a mean distribution similarity of $K^{xy} = 0.91$ (see Figs. 29–31). Moreover, we even obtained the same mean and standard deviation as the real data for purchase intent across all surveys $\text{E}[\text{PI}] = 4.0 \pm 0.1$. At the same time, for the best reference set, correlation attainment only reached $\rho = 50\%$ compared to $\rho = 92\%$ for Gem-2f SCs prompted with demographic markers, which suggests that if LLMs are not prompted with a detailed enough persona they will rate products

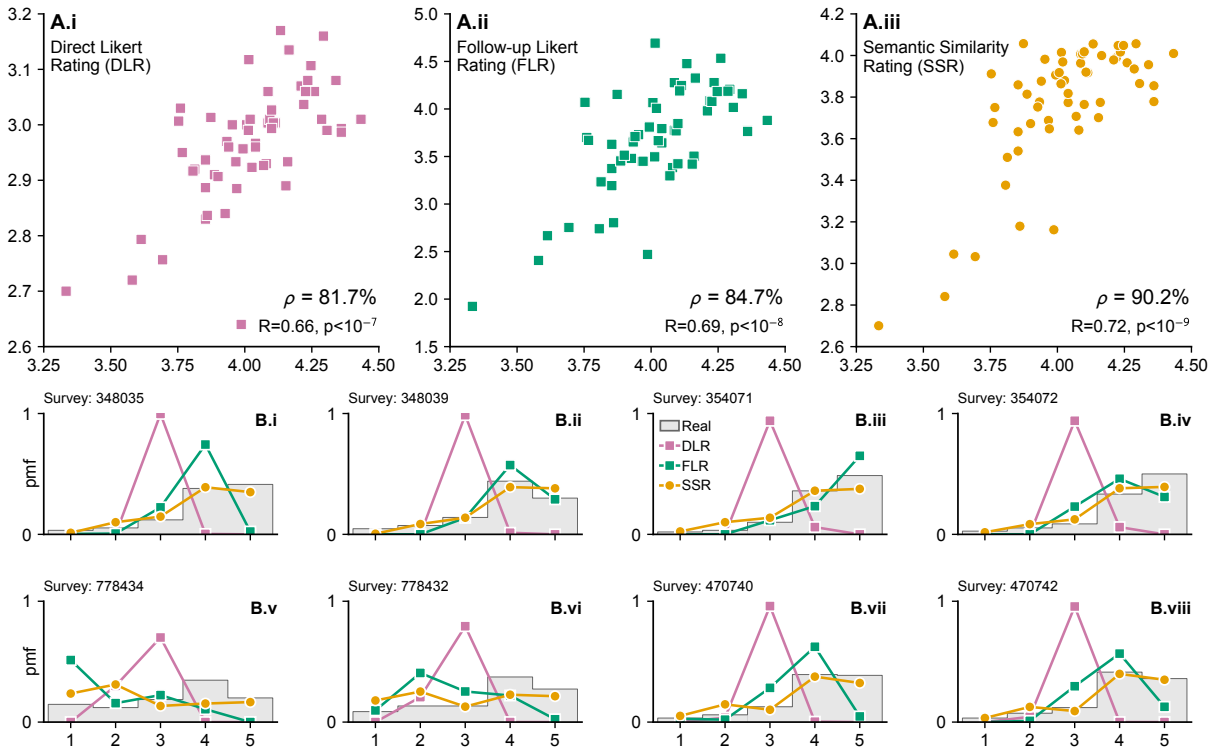

Figure 2: Comparison of real and synthetic surveys based on GPT-4o with $T_{\text{LLM}} = 0.5$. (A) Mean purchase intent comparison for (A.i) Direct likert ratings (DLRs), (A.ii) textual elicitation with follow-up Likert ratings (FLRs) and (A.iii) semantic similarity ratings (SSRs). (B) Eight example survey response distributions for real surveys and the corresponding synthetic surveys based on DLR, FLR, and SSR, respectively.

more positively in general and will not leverage the actual information in the product concepts enough to produce a meaningful signal. We obtained similar results in further experiments for both models (see Figs. 23–28).

4.4 Additional Results

While the SSR method is of quantitative nature, the underlying free-text responses make it possible to obtain qualitative feedback on product concepts. Comparing textual responses by humans to those generated by LLMs, we find that the latter are far richer in information, highlight positive features, and raise explicit concerns about less likable product properties. SCs may thereby enrich product research beyond quantitative analyses (see App. E).

To test how the SSR method would perform for indicators other than purchase intent, we ran a single experiment for the question "How relevant was the concept?" which was posed to the same human participants within each survey. Taking the average over three new reference sets that were constructed as Likert-scale anchors for this question, we found that the synthetic responses by Gem-2f achieved a correlation attainment $\rho = 82\%$ for SSR and $\rho = 91\%$ for FLRs (cf. Fig. 35). At the same time, the synthetic survey distributions achieved similarity values of $K^{xy} = 0.81$ for SSR $K^{xy} = 0.62$ for FLRs (cf. Figs. 33-34), demonstrating the method's potential for generalization.

We further wanted to test how much information the LLMs actually leverage from the product concept beyond coarse-grained features such as demographics and product properties. To this end, we trained 300 LightGBM classifiers on one random half of the studies each and analyzed predicted responses for the other half (see App. D). LightGBM, despite being trained on in-sample data, achieved a correlation attainment of only $\rho = 65\%$, compared to $\rho = 83\%$ for FLRs and $\rho = 88\%$ for SSRs. Regarding distributional similarity, LightGBM outperformed FLRs with $K^{xy} = 0.80$ versus $K^{xy} = 0.72$. However, SSR distributions were still closer to the real survey results with $K^{xy} = 0.88$, highlighting that zero-shot LLM elicitation—which requires no access to training data from the surveys—synthesizes human responses more effectively than a supervised ML model that does.

To see how a more parsimonious setup would perform, we repeated most of the experiments and replaced the image stimulus with a text stimulus that contained only the product description. We find that doing so mildly reduces the performance for most experiments. Success metrics for all experiments can be found in Tab. 1.

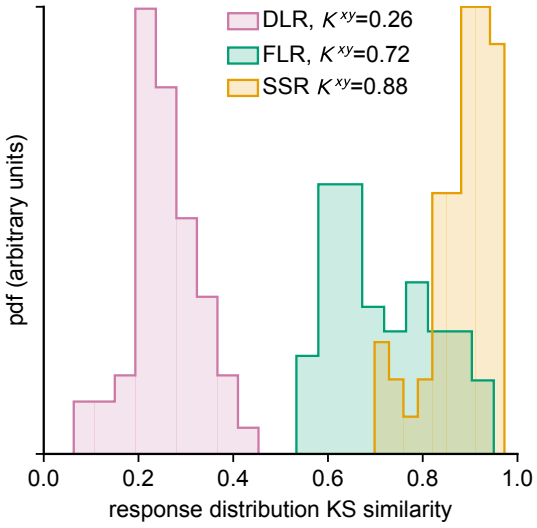

Figure 3: Comparison of purchase intent distribution similarity between real and synthetic surveys based on GPT-4o with $T_{\text{LLM}} = 0.5$ for direct Likert ratings (DLRs), textual elicitation with follow-up Likert ratings (FLRs) and semantic similarity ratings (SSRs).

5 Discussion and Conclusion

Our results show that LLM-based synthetic consumers can reproduce core outcomes of traditional consumer concept testing with surprising fidelity. In particular, the semantic similarity rating (SSR) approach yields both realistic distributions of Likert responses and robust product rankings that attain over 90% of the maximum correlation with human data, based on test–retest reliability. These findings suggest that many of the shortcomings of prior attempts at using LLMs as survey respondents—such as skewed distributions, over-positivity, or regression-to-the-mean—are not intrinsic limitations of LLMs, but rather artifacts of how responses were elicited. By shifting from direct elicitation of Likert responses to textual elicitation and SSR, we resolve many of these artifacts and unlock richer, more interpretable data.

Importantly, no training data or fine-tuning on consumer responses was required. This makes the method widely applicable and inexpensive compared to training or calibration-heavy alternatives. The SSR approach functions as a plug-and-play tool: it translates free-text responses into Likert distributions in a transparent, interpretable way, preserving comparability with canonic survey-based consumer research while also capturing the nuance of unconstrained responses.

A key advantage of this approach is the retention of qualitative information. Whereas human Likert ratings are often accompanied by minimal free-text justifications, LLM-based synthetic consumers readily provide detailed rationales that explain why a product might be attractive or unattractive. These rationales can be mined for themes, objections, or value propositions in ways that raw Likert

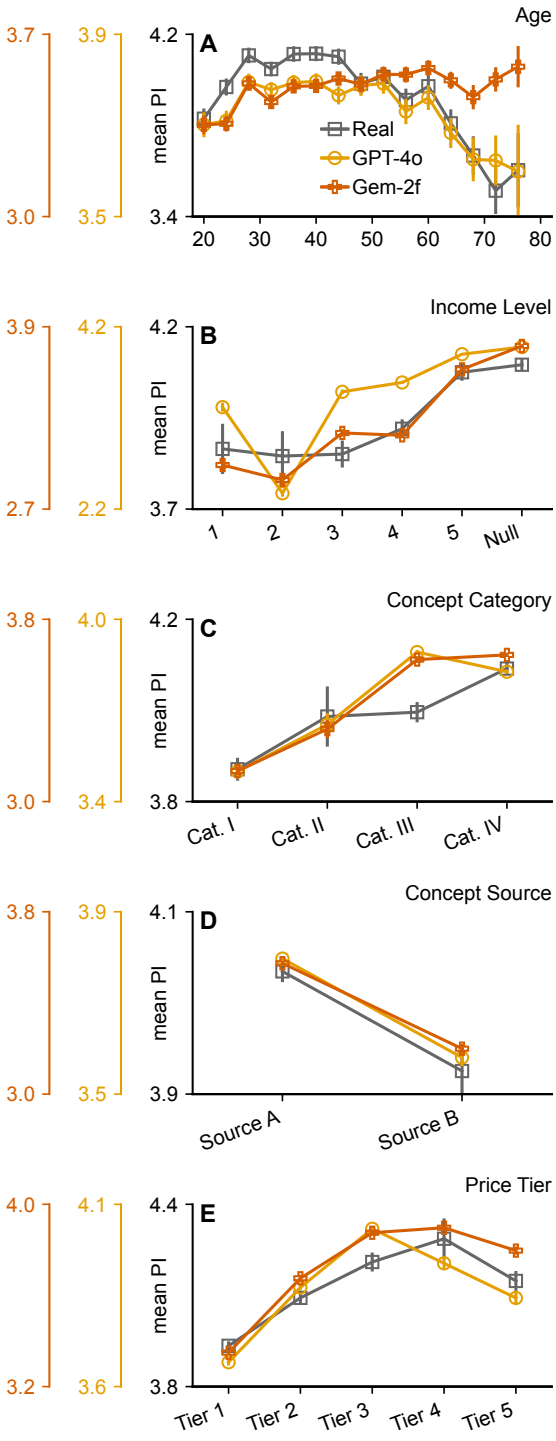

Figure 4: Mean purchase intent stratified by five demographic and product features (shown are results from the SSR method for both GPT-4o and Gem-2f). Error bars represent standard errors.

scores cannot. Moreover, we observe that synthetic responses appear less prone to the positivity bias common in human surveys, producing a wider spread of purchase intent. This broader dynamic range may provide companies with more discriminative signals when evaluating early-stage concepts.

While promising, the method is not without limitations. SSR relies on carefully designed reference statements, and our results show that different anchor sets can lead to slightly different mappings. Averaging across sets mitigates this issue, but future work could optimize reference statements iteratively, or even generate them dynamically with LLMs to maximize alignment with human data. Note that the reference sets created herein were manually optimized for the 57 surveys subject to this study, which means it remains elusive how well they would perform for other surveys. Another limitation concerns demographics: although LLMs captured some demographic patterns (e.g., age and income) quite well, others (e.g., gender, region, ethnicity) were not consistently replicated. This suggests that while persona conditioning does influence synthetic responses, it cannot yet be treated as a reliable proxy for all subpopulations. Researchers must therefore use caution when interpreting subgroup analyses from synthetic panels.

More fundamentally, the usefulness of SSR is bounded by the knowledge domains represented in the LLM's training data. LLMs are known to hallucinate when asked about unfamiliar topics, and SSR does not eliminate this risk. The reason our approach succeeds in oral care products, and even reflects demographic conditioning, is likely that the model has been exposed to abundant human discussions of these categories in its training corpus (e.g., online forums and consumer reviews). For domains where such background knowledge is sparse or absent, SSR will not conjure valid consumer preferences. Thus, it is important not to view synthetic surveys as universally reliable, but rather as tools whose validity depends on the alignment between training data and the survey domain.

Additionally, SSR's performance depends on the choice of embedding model and similarity measure. Although cosine similarity proved effective, further benchmarking could reveal alternative embedding spaces (e.g., domain-specific encoders) that yield stronger alignment. Finally, while synthetic consumers reproduce human-like distributions and rankings, they cannot fully capture the real-world contingencies of purchasing behavior, such as budget constraints, cultural context, or marketing exposure.

There are several avenues for extending this work. First, the method can be generalized to survey questions beyond purchase intent. By designing reference sets for other Likert constructs such as relevance, satisfaction, or trust we may extend the approach to larger surveys or other applications. Second, optimization strategies could improve SSR: parameters quantifying how a single response is mapped to a distribution could be tuned automatically to maximize correlation with held-out human data. Third, more sophisticated prompting strategies could be explored, such as asking LLMs to generate long free-text responses that are then summarized before mapping to Likert anchors of different questions at once, or multi-stage pipelines in which one LLM generates responses and another critiques or calibrates them. While such methods may be more computationally expensive, they could improve both interpretability and alignment.

Finally, there is an open question about combining SSR with light fine-tuning approaches. Although we deliberately avoided training data here to demonstrate generality, hybrid methods where SSR is used in tandem with calibration or prompt optimization may achieve even higher fidelity. Crucially, however, SSR provides a low-cost baseline that narrows the gap between synthetic and human survey data without requiring retraining.

If further validated, SSR-enabled synthetic consumers could substantially change how early-stage product research is conducted. Instead of commissioning large human surveys for every product idea, companies could first screen concepts synthetically, reserving human panels for the most promising candidates. This would reduce costs, accelerate iteration, and enable smaller firms to access consumer insights that were previously out of reach. At the same time, the availability of detailed synthetic rationales could complement human panels, offering a richer understanding of consumer perceptions.

In summary, by combining interpretability, statistical reliability, and qualitative richness, SSR addresses many of the challenges that have constrained the use of LLMs as synthetic survey respondents. While not a wholesale replacement for human research, SSR establishes a credible framework for augmenting and accelerating consumer insight generation.

Acknowledgments

The original manuscript was written by all authors, ChatGPT-5 was subsequently used in all sections to reformulate and condense the text.

References

[1] Kevin Lane Keller and Philip Kotler. *Marketing Management.* Pearson Education, 15th edition, 2015.

[2] Karl T Ulrich and Steven D Eppinger. *Product Design and Development.* McGraw-Hill Education, 6th edition, 2015.

[3] ESOMAR. Global market research 2024, 2025.

[4] Kevin J Clancy and Peter C Krieg. *Market New Products Successfully: Using Simulated Test Markets for Accurate Forecasting.* Lexington Books, 2001.

[5] Alvin J Silk and Glen L Urban. Pre-test market evaluation of new packaged goods: A model and measurement methodology. *Journal of Marketing Research*, 13(2):171–191, 1976.

[6] Linda F Jamieson and Frank M Bass. Purchase intentions and product purchase probability. *Journal of Marketing Research*, 26(3):336–345, 1989.

[7] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55, 1932.

[8] Jon A Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236, 1991.

[9] Jon A Krosnick. Survey research. *Annual Review of Psychology*, 50(1):537–567, 1999.

[10] Peiyao Li, Vineet Kumar, Donald Ngwe, and Mengjie Sun. Determining the validity of large language models for automated perceptual analysis. *Marketing Science*, 43(2):254–266, 2024. Frontiers.

[11] James Brand, Ayelet Israeli, and Donald Ngwe. Using LLMs for market research. Technical report, Harvard Business School Working Paper 23-062, 2024.

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

[13] James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, Jennifer Larson, R. Blake Patterson, and Matthew Carlson. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4):401–416, 2024.

[14] Christina Kaiser, Johannes Kaiser, Valentina Manewitsch, et al. Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25)*, 2025.

[15] Aadesh Salecha, Molly E. Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H. Ungar, and Johannes C. Eichstaedt. Large language models show human-like social desirability biases in survey responses. *arXiv:2405.06058*, 2024.

[16] Wenpeng Yin, Jennifer Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation, and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3914–3923. Association for Computational Linguistics, 2019.

[17] Gary King, Christopher J.L. Murray, Joshua A. Salomon, and Ajay Tandon. Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, 98(1):191–207, 2004.

[18] Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning (PMLR Volume 202)*, pages 337–371. PMLR, 2023.

[19] Suhyun Cho, Jaeyun Kim, and Jang Hyun Kim. LLM-based doppelgänger models: Leveraging synthetic data for human-like responses in survey simulations. *IEEE Access*, 12:178917–178927, 2024.

[20] Bernard J Jansen, Soon-Gyo Jung, and Joni Salminen. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020, 2023.

[21] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157. Curran Associates Inc., 2017.

Appendix

A Detailed Methods

A.1 Data

We base our study on a corpus of 57 consumer research surveys, conducted using a digital consumer research platform, supplied by a leading personal care corporation. Each survey relates to a unique hypothetical personal care product concept designed for the US market. In the dataset, a product concept is represented on a presentation slide that contains at minimum a text description and in many instances a concept image, as well. Every survey had a unique set of participants, with sizes ranging from $N_s = 150$ to $N_s = 400$. In total, the corpus has 9,300 unique participants. For most of the surveys, age, gender, and location of the participants is known. To a lesser extent, income level is reported, and only nine surveys contain information on consumer ethnicity.

While surveys prompt participants to score product concepts along various dimensions, we focus on the central question of *purchase intent*, often phrased as "Based on everything you've seen and heard, how likely are you to purchase the product?". This response was scored on a 5pt Likert scale, requiring participants to select one of the integer numbers $i = 1, 2, 3, 4, 5$.

A.2 Definitions

Let the corpus of all surveys be called $\mathcal{S}$. A single survey $s \in \mathcal{S}$ consists of:

(1) $N_s = |C_s|$ consumers $c \in C_s$ with demographic attributes $\boldsymbol{d}_c = \{d_{c,1}, d_{c,2}, \ldots, d_{c,D}\}$ containing features such as age, gender, income, location, and ethnicity, as well as
(2) a single product concept.

Note that we need not formally distinguish between product concepts and surveys as each survey only relates to a single unique concept, hence s may either denote a survey or a concept depending on the context.

Let r_c be the Likert scale rating response provided by a human consumer c after reviewing a product concept s, and asked about their purchase intent (since there are no consumers that repeat across surveys, we need not index the response by s). Response r_c may be equal to any of the integer numbers $i = 1, 2, 3, 4, 5$. Having produced a Likert scale response, each consumer $c \in C_s$ of a given survey s is associated with a response probability mass function (pmf) of $p_c(i) = \delta_{ir_c}$, where δ_{ir_c} is the Kronecker delta function. The whole survey distribution, which aggregates responses from all consumers, is then given by:

$$p_s(i) = \frac{1}{N_s} \sum_{c \in C_s} \delta_{ir_c} \tag{1}$$

The mean purchase intent of the concept s is then calculated as:

$$\mathrm{PI}_s = \sum_{i=1}^{5} i p_s(i). \tag{2}$$

We describe as a *synthetic consumer* $\tilde{c}$ an LLM that was prompted to impersonate a human consumer with demographic attributes $\boldsymbol{d}_c$ or a subset thereof, see Section A.4.

All of the definitions outlined above apply for synthetic consumers $\tilde{c}$, as well, however with the important distinction that we do not restrict those to reply with single-response distributions, i.e. we do not require that the response pmf is a Kronecker delta function. Instead, a synthetic consumer response may yield an arbitrary pmf $p_{\tilde{c}}(i)$ on the Likert scale. As we shall see in Section A.4, this involves mapping a textual response $t_{\tilde{c}}$ to a Likert scale rating $r_{\tilde{c}}$ integer, or a pmf $p_{\tilde{c}}(i)$.

Henceforth, we will denote real data with the superscript x and synthetic data with the superscript y.

A.3 Success Metrics

We define two success metrics, one to measure the distributional similarity between outcomes of synthetic and real surveys, and another to measure the degree to which synthetic consumers replicate the concept ranking obtained from real surveys.

Distributional Similarity. Synthetic consumer panels should replicate real consumer purchase intent distributions as accurately as possible. To this end, we define the distributional similarity between a synthetic and real survey s as the complement of the Kolmogorov-Smirnov (KS) distance:

$$\text{KS sim}_s^{xy} = 1 - \sup_{r=1,\ldots,5} |F_s^x(r) - F_s^y(r)|. \tag{3}$$

Likert data responses are ordinal and because there is no measurable concept of distance between the integer responses, technically, the (KS) distance is an inappropriate measure. However, we find that in practice KS distance has various advantages, for instance it is simple to interpret, as the maximum distance between two CDFs which is always a number between 0 (where distributions are equal) and 1 (where distributions are entirely dissimilar). Second, the ordinality of the data is respected, i.e. it does make a strong difference whether two distributions have peaks that lie close or far away from each other.

At times, we will compare KS similarity to distributional cosine similarity defined as

$$C_s^{xy} = \frac{\boldsymbol{p}_s^x \cdot \boldsymbol{p}_s^y}{|\boldsymbol{p}_s^x||\boldsymbol{p}_s^y|} \tag{4}$$

which does not respect the scale's ordinality. Here, we treat the pmf as a vector $\boldsymbol{p} = (p(1), \ldots, p(5))$.

We denote as $K^{xy} = \mathrm{E}[\text{KS sim}_s^{xy}]$ the mean distributional similarity over all surveys (and analogously, $C^{xy} = \mathrm{E}[C_s^{xy}]$).

Concept Ranking Similarity and Correlation Attainment. A concept's popularity in terms of mean purchase intent should rank similarly for both synthetic surveys as well as real surveys. To measure how similarly concepts are perceived, we compute the Pearson correlation between the mean purchase intents from synthetic and real surveys PI^y and PI^x, respectively:

$$R^{xy} = \text{corr}[\text{PI}^x, \text{PI}^y]. \tag{5}$$

Naively, we should thus expect perfect synthetic consumers to yield R^{xy} close to 1. However, since we observe that the mean purchase intents of real surveys lie relatively close to each other with $\mathrm{E}[\text{PI}^x] = 4.0$ and $\text{Std}[\text{PI}^x] = 0.2$, we must consider the possibility that these results are influenced by noise and hidden biases in a non-negligible manner. We therefore ask: were the surveys repeated with a new cohort of similarly drafted real consumers, how well would the mean purchase intent of the repeated surveys correlate with the mean purchase intent of the original surveys? This value, R^{xx}, should be the theoretical maximum concept ranking similarity we judge the performance of the synthetic consumers by. In other words, a high concept ranking correlation between synthetic and real consumers is one that approaches the retest correlation of real consumer responses.

Although we cannot obtain a traditional test–retest reliability estimate by repeating each survey with a new cohort of real consumers of size N_s, we can simulate retest scenarios a large number of times where we randomly split survey participants into test and control cohorts of size $N_s/2$.

To obtain a reliability measure of the concept ranking, we perform the following comparison: For every survey s, we split the participant set C_s in half at random. One half $C_{s,t}$ will be called the *test cohort*, whose responses constitute the central survey results. We call the second half $C_{s,c}$ the *control cohort*, whose responses represent a repeated survey to control and compare the results of the first survey. Then, for the corresponding synthetic survey with participants C_s^y, we randomly sample a test cohort of the same size as the corresponding real test and control cohorts, respectively. We follow this cohort construction procedure once for every survey, achieving corresponding average purchase intents $\text{PI}_{s,t}^x$, $\text{PI}_{s,c}^x$, and $\text{PI}_{s,t}^y$, such that correlation coefficients $R^{xx} = \text{corr}\left[\text{PI}_{s,t}^x, \text{PI}_{s,c}^x\right]$ and $R^{xy} = \text{corr}\left[\text{PI}_{s,t}^x, \text{PI}_{s,t}^y\right]$ can be computed. Repeating this experiment $m = 2\,000$ times and taking the average over the respective correlation coefficient, we obtain *correlation attainment*

$$\rho = \frac{\mathrm{E}[R^{xy}]}{\mathrm{E}[R^{xx}]}, \tag{6}$$

quantifying how close the correlation coefficient between real and synthetic consumers is to the theoretical maximum.

A.4 Synthetic Response Generation

For every human consumer c, a *synthetic consumer* $\tilde{c}$ was constructed by priming an LLM to be a participant in a product research survey and to impersonate a consumer c with the same or a subset of demographic attributes $\boldsymbol{d}_{\tilde{c}} \subseteq \boldsymbol{d}_c$. Then the LLM was shown the product concept, in the form of an image containing the text and potentially an image of the product (see App. B). Subsequently, the LLM was prompted with the question "How likely are you to purchase the product?", and a response was sampled. Due to the stochastic nature of LLMs, we designed our experimental setup such that any number of repeat samples n could be drawn upon submitting each prompt, enabling us to average results over multiple samples. In the following analysis we use $n = 2$ samples, which we found was sufficient to obtain stable results.

We focused on models by Google and OpenAI; after initial experiments with different models such as "'gemini-1.5-flash," "gemini-2.5", and "o3" we settled on "gemini-2.0-flash" (referred to as "Gem-2f" throughout the text) and "gpt-4o" ("GPT-4o") for production runs as these models gave the most consistent responses across all experiment types. Experiments were run with parameters $p_{\text{top}} = 0.9$ and temperature $T_{\text{LLM}} \in \{0.5, 1.5\}$ if not noted otherwise.

Below we describe different approaches to generate synthetic responses.

A.4.1 Direct Likert Rating. The simplest approach to generate a Likert scale rating from an LLM which has been presented with a product concept, is to treat the LLM like a human participant and let it respond with a single token, i.e. one of the integer responses 1, 2, 3, 4, or 5. This approach is straightforward and produces results with minimal effort.

A.4.2 Follow-up Likert Rating (Textual Elicitation Before Rating). A slightly more sophisticated approach is to let the LLM first express its liking of the product concept in a brief but otherwise unconstrained text response $t_{\tilde{c}}$, and only afterward let it summarize this

in a single integer response $r_{\tilde{c}}$. Specifically, after priming the LLM with its demography and showing a product concept, we prompt it with the question "How likely are you to purchase the product?", and stating "Reply briefly to any questions posed to you." in the system prompt. After sampling the response, we prompt the LLM to be a "Likert rating expert" and request it to map the text response it just gave, to the corresponding integer response on a 5pt Likert scale. The same LLM that generates a response is also tasked with rating it, using $p_{\text{top}} = 1$ and $T_{\text{LLM}} = 0.3$.

A.4.3 Semantic Similarity Rating (SSR). Mapping textual responses to Likert scale ratings is, however, non-trivial, as a response rarely maps exactly to one and only one rating. For instance, a response may read "I'd probably buy it. I like that it's easy to use and I can take it with me. Plus, the price isn't too bad." Depending on how the scale is defined and who is on the receiving end of this statement, the response most likely would map to a "4" or a "5". The statement could be easily interpreted as purchase being very likely (5) when imagining a casual conversation about purchase intent. Others might interpret this response as just "likely" (4). Hence, there's an inherent ambiguity in the textual responses that gets discounted through the mapping onto a single number.

We therefore propose an alternative procedure that maps a textual response to a distribution of Likert scale ratings. To do this, we construct reference statements σ_r that each map to a Likert scale rating, then estimate the similarity of the response text $t_{\tilde{c}}$ to each of these reference statements, and use the similarities to construct a pmf of Likert scale ratings. In practice, we construct m such reference statement sets Σ_i where $i = 0, \ldots, m-1$, across which we eventually average the respective single-response pmfs to obtain a single-response result pmf. In this analysis, we use $m = 6$ sets (see App. C.1). They are all similar but not identical, and designed to capture the different ways a consumer may express their purchase intent. However, as this use of multiple sets of reference statements is more an implementation detail than a theoretical innovation, and since it will complicate the mathematical notation unnecessarily, we do not explicitly mention it in the following.

To effectively compute the similarity between a response text $t_{\tilde{c}}$ and a reference statement σ_r, we retrieve embedding vectors $\boldsymbol{v}_{\sigma_r}$ and $\boldsymbol{v}_{t_{\tilde{c}}}$ from a text-embedding model for each of the reference statements as well as the response text from synthetic consumer $\tilde{c}$, respectively. In this study, we exclusively use OpenAI's "text-embedding-3-small" model. Tests with "text-embedding-3-large" left the results virtually unchanged. With these vectors, every response can be mapped to a similarity in the embedding space by means of cosine similarity as

$$\gamma(\sigma_r, t_{\tilde{c}}) = \frac{\boldsymbol{v}_{\sigma_r} \cdot \boldsymbol{v}_{t_{\tilde{c}}}}{|\boldsymbol{v}_{\sigma_r}||\boldsymbol{v}_{t_{\tilde{c}}}|}. \tag{7}$$

We then interpret this similarity as proportional to a response probability p_r for integer response r, such that

$$p_{\tilde{c},i}(r) \propto \gamma(\sigma_{r,i}, t_{\tilde{c}}) - \gamma(\sigma_{\ell,i}, t_{\tilde{c}}) + \epsilon\delta_{\ell,r} \tag{8}$$

where ℓ is the reference statement with minimum similarity over the corresponding set Σ_i and normalization $\sum_{r=1}^{5} p_{\tilde{c},i}(r) = 1$. Note that subtracting the minimum similarity over the reference statement set $\gamma(\sigma_{\ell,i}, t_{\tilde{c}})$ is a way to adjust for potential low variance in the similarity scores. In practice, we observe that within the space of all embeddable language, the numerical difference between extremes like "It's very unlikely that I'd buy it." and "It's very likely that I'd buy it." is numerically small, and so if we do not subtract the minimum similarity, the resulting pmf will be almost entirely flat. For $\epsilon = 0$, this equation can be read as follows: For every similarity, subtract the minimum similarity over the reference statement set. Normalize the remaining similarities by the total sum to obtain a probability mass function (pmf). Of course, following this procedure means that for every textual response we obtain a rating distribution where one of the ratings has zero likelihood to occur. The parameter ϵ offsets this bias and makes it more controllable.

To make this mapping from a similarity to a probability mass function (pmf) even more controllable, we can introduce a temperature-like parameter that controls how "smeared out" the resulting pmf should be

$$p_{\tilde{c},i}(r, T) \propto p_{\tilde{c},i}(r)^{1/T} \tag{9}$$

with $\sum_r p_{\tilde{c},i}(r, T) = 1$.

While we restrict our study to $\epsilon = 0$ and $T = 1$, it is worth introducing them as levers to make the SSR mapping more controllable. One can, for instance, design an optimization procedure to find values for ϵ and T that yield the best SSR mapping in terms of the success metrics defined in Section A.3. A first test suggests that $T = 1$ is a reasonable choice as a rule of thumb, but that there is optimization potential: clearly there are optima to be found around $T = 1$ both for maximizing correlation and distribution similarity (see Fig. 32). A full Python code implementation of SSR is given in Appendix C.2.

B Product Concept Example

Fig. 5 shows an examples of a product concept image similar to those used in the study. When we refer to "image stimulus" in the main text, an image like this, including either both an illustration and the concept description or only a concept description was supplied to an LLM synthetic consumer. When we refer to "text stimulus," only the text was given to the synthetic consumer. For the latter, the text was previously transcribed from the product concept image using GPT-4o.

C Additional Material for SSR

C.1 Reference Statement Sets

To map free-text responses onto a 5pt Likert scale, we constructed six sets of anchor statements, five statements in each set (one for each Likert category 1 through 5). These anchors serve as semantic prototypes against which model-generated responses are compared in embedding space. Each anchor statement was written to reflect the intensity of purchase intent associated with its corresponding Likert rating:

- The lowest anchor expresses a purchase to be unlikely.
- The middle anchor reflects indifference or uncertainty.
- The highest anchor conveys intent or possible intent to purchase.

The remaining two intermediate anchors were formulated as semantically in between their adjacent statements. The anchors were designed to be short, generic, and domain-independent, such that they could plausibly apply to any consumer product concept. Their

Figure 5: A surrogate product concept similar to those used in the 57 concept surveys.

purpose is not to capture the nuances of any specific product, but to provide a reference gradient of intent from "purchase improbable" to "purchase probable." This approach allows defining the Likert scale in a way that adapts to the stylistic tendencies of LLM responses, ensuring that the full range of intent is captured.

C.2 SSR Implementation

We give a full Python implementation of the SSR method, available on GitHub at https://github.com/pymc-labs/semantic-similarity-rating. The package offers a simple programming interface for generating SSR-based Likert scale response distributions from LLM responses.

D Machine-Learning Comparison Based on LightGBM

To benchmark the performance of zero-shot LLM elicitation against a classical machine learning approach, we trained gradient-boosted decision tree models (LightGBM [21]) on subsets of the survey data. The goal was to assess how well a model trained on demographic and product features could reproduce individual Likert ratings compared to synthetic responses generated by LLMs.

We considered the complete set of 57 consumer concept surveys used in the main study. For each of 300 iterations, we randomly split the surveys in half. One half (28 surveys) was used for training, the other half (29 surveys) for testing. This setup mirrors the cross-study generalization scenario relevant for real-world applications, where new concepts and respondents are unseen during training.

On each training split, we fitted a LightGBM classifier with the following input features:

- **Demographics (5 features):** age, gender, income tier, region, and ethnicity.
- **Concept attributes (3 features):** category of personal care, price tier, and concept source.

The target variable was the purchase-intent rating on a 5-point Likert scale. Models were trained with default LightGBM hyperparameters, using multiclass classification (with the five Likert values treated as separate classes). Missing feature data was assigned a "Null" category.

For each held-out study, we predicted Likert responses for all respondents and aggregated them into synthetic response distributions. We then compared these predictions to the original human survey data using two metrics: (i) test–retest reliability (ρ), defined analogously to the procedure in the main text, i.e. by Pearson correlation between synthetic and human mean purchase intents of the 29 surveys of the test set, split in half once again, and (ii) distributional similarity (K^{xy}), defined as the complement of the

Kolmogorov–Smirnov distance between the predicted and observed Likert distributions for the 29 surveys in the test set.

Across 300 iterations, LightGBM achieved a correlation attainment of $\rho = (64.6 \pm 1.0)\%$, substantially below both follow-up Likert ratings ($83.2 \pm 0.7\%$) and SSR-based elicitation with GPT-4o at $T_{\text{LLM}} = 0.5$ and image stimulus ($88.3 \pm 0.7\%$). For distributional similarity, LightGBM ($K^{xy} = 0.797 \pm 0.002$) outperformed follow-up Likert ratings (0.716 ± 0.001) but was surpassed by SSR (0.883 ± 0.001).

This analysis shows that a simple supervised ML model trained on demographic and product features cannot match the fidelity of zero-shot LLM elicitation in recovering human-like response behavior. While LightGBM achieves moderate distributional alignment, its markedly lower reliability underscores the advantage of LLM-based methods that leverage semantic understanding of product descriptions without requiring additional training data.

E Textual Responses Allow for Qualitative Evaluations

As a byproduct of applying the SSR method to obtain quantitative ratings, the textual responses generated by LLMs are rich in information and make it possible to evaluate the product concept in more detail.

While the real surveys focused on ratings on a Likert scale, they also asked people to write free text as responses to the questions "What did/didn't you like about the concept?" The replies to these open questions are lacking depth and only seldom provide important feedback. Typically, they are rather short like "It's good", or just repeat information contained in the concept like "Not much, just the steps and how it tells you what it was for." Less frequently, participants gave actual feedback about what they liked, for instance one such response reads "Inexpensive and affordable. New & light. [Application] from your own home." In contrast, the replies about purchase intent by synthetic consumers provide much more in-depth feedback about why or why not they would possibly purchase the product. For instance, one synthetic consumer wrote: "The ease of use and [...] safety are appealing, but I'd want to know more about its effectiveness and any potential side effects." Another responded: "The ease of use and the promise of no [...] sensitivity are appealing. Plus, it's from a trusted brand."

Similarly, synthetic consumers rarely held back with their criticisms, which, at times, came written in the style of the persona they were asked to imitate. For a less positively received concept, GPT-4o synthetic consumers responded "It seems a bit too high-end for my needs and budget" and "[These body parts] don't usually bother me, so I don't think I need it" while two based on Gem-2f wrote e.g., "Seems kinda bougie for [this kind of product]" and "Sounds expensive, and I'm not sure I buy all that 'microbiome' talk. I'll stick with what I know", respectively.

In total, the responses generated by LLMs can be leveraged to obtain detailed feedback on why or why not a product concept was rated with higher or lower purchase intent. Additionally, synthetic consumers seem to suffer less from a positivity bias, as demonstrated by the wider spread of mean purchase intent measured in the previous section as well as confirmed by the qualitative analysis of synthetic responses.

Received October 14, 2025

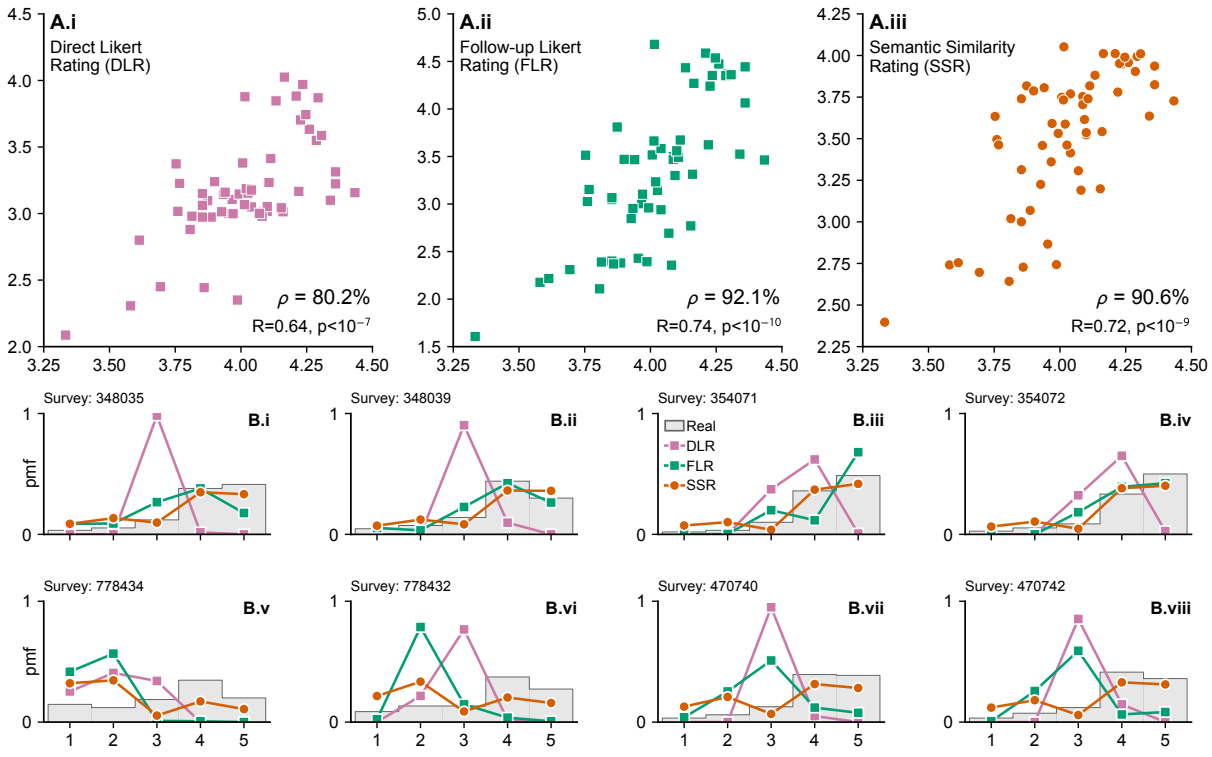

Figure 6: Comparison of real and synthetic surveys based on Gem-2f with $T_{\text{LLM}} = 0.5$. (A) Mean purchase intent comparison for (A.i) Direct likert ratings (DLRs), (A.ii) textual elicitation with follow-up Likert ratings (FLRs) and (A.iii) semantic similarity ratings (SSRs). (B) Eight example survey response distributions for real surveys and the corresponding synthetic surveys based on DLR, FLR, and SSR, respectively.

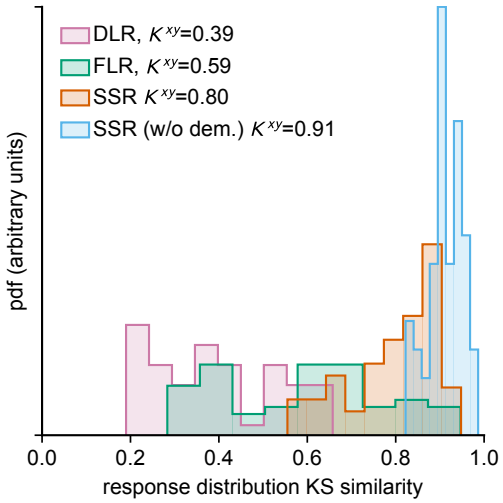

Figure 7: Comparison of purchase intent distribution similarity between real and synthetic surveys based on Gem-2f with $T_{\text{LLM}} = 0.5$ for direct Likert ratings (DLRs), textual elicitation with follow-up Likert ratings (FLRs), semantic similarity ratings (SSRs), and best-set SSRs for an experiment where synthetic consumers where prompted without demographic markers.

Table 1: Metrics for all experiments on purchase intent. "Direct" refers to direct Likert rating elicitation (DLR), "Textual" refers to free-text responses followed by SSR and FLR. We show correlation attainment ρ, distributional similarities K^{xy} and C^{xy}, mean purchase intent correlation R^{xy} between human and synthetic surveys, including results for best-comparison set Σ (see App. C.1). We also report mean survey purchase intent and its standard deviation.

Elicit.	Dem.	Model	Stim.	T_{LLM}	Best Σ	ρ (%)		K^{xy}		R^{xy}		C^{xy}		$\text{E}[\text{PI}_s] \pm \text{std}$	
						SSR	Lik.	SSR	Lik.	SSR	Lik.	SSR	Lik.	SSR	Lik.
Direct	Full	GPT-4o	Text	1.5			88.5		0.37		0.717		0.39		2.95 ± 0.44
Direct	Full	GPT-4o	Text	0.5			89.6		0.36		0.720		0.38		2.96 ± 0.45
Direct	Full	GPT-4o	Image	1.5			78.7		0.29		0.648		0.29		2.97 ± 0.16
Direct	Full	GPT-4o	Image	0.5			81.7		0.26		0.661		0.26		2.96 ± 0.11
Direct	Full	Gem-2f	Text	1.5			68.4		0.46		0.546		0.48		3.28 ± 0.50
Direct	Full	Gem-2f	Text	0.5			67.5		0.45		0.541		0.47		3.28 ± 0.50
Direct	Full	Gem-2f	Image	1.5			82.5		0.41		0.660		0.41		3.17 ± 0.40
Direct	Full	Gem-2f	Image	0.5			80.2		0.39		0.640		0.40		3.17 ± 0.40
Textual	Full	GPT-4o	Text	1.5	4	85.0		0.85		0.691		0.94		3.75 ± 0.40	
Textual	Full	GPT-4o	Text	0.5	4	83.1		0.84		0.680		0.92		3.71 ± 0.45	
Textual	Full	GPT-4o	Text	0.5	avg.	82.5	69.2	0.84	0.64	0.675	0.562	0.94	0.70	3.63 ± 0.42	3.39 ± 0.79
Textual	Full	GPT-4o	Image	0.5	1	91.9		0.82		0.740		0.93		3.67 ± 0.36	
Textual	Full	GPT-4o	Image	0.5	avg.	90.2	84.7	0.88	0.72	0.724	0.687	0.96	0.80	3.77 ± 0.31	3.67 ± 0.55
Textual	Full	Gem-2f	Text	1.5	3	76.3		0.81		0.611		0.92		3.56 ± 0.40	
Textual	Full	Gem-2f	Text	0.5	3	72.7		0.81		0.581		0.91		3.56 ± 0.42	
Textual	Full	Gem-2f	Text	0.5	avg.	73.0	73.5	0.80	0.62	0.583	0.589	0.91	0.69	3.52 ± 0.45	3.57 ± 0.86
Textual	Full	Gem-2f	Image	0.5	5	92.4		0.82		0.737		0.92		3.64 ± 0.45	
Textual	Full	Gem-2f	Image	0.5	avg.	90.6	92.1	0.80	0.59	0.720	0.736	0.91	0.64	3.51 ± 0.42	3.33 ± 0.75
Textual	None	GPT-4o	Text	1.5	4	50.1		0.92		0.409		0.98		4.05 ± 0.12	
Textual	None	GPT-4o	Text	0.5	4	45.3		0.91		0.390		0.97		4.09 ± 0.16	
Textual	None	GPT-4o	Text	0.5	avg.	47.4	41.2	0.91	0.57	0.408	0.324	0.98	0.77	3.92 ± 0.14	4.61 ± 0.51
Textual	None	Gem-2f	Text	1.5	3	60.5		0.87		0.481		0.95		3.85 ± 0.24	
Textual	None	Gem-2f	Text	0.5	3	49.5		0.87		0.411		0.95		3.90 ± 0.26	
Textual	None	Gem-2f	Text	0.5	avg.	57.4	58.0	0.91	0.62	0.481	0.480	0.97	0.75	3.90 ± 0.26	4.26 ± 0.58
Textual	None	Gem-2f	Image	0.5	4	50.1		0.91		0.414		0.98		4.09 ± 0.07	
Textual	None	Gem-2f	Image	0.5	avg.	15.5	64.3	0.91	0.67	0.143	0.530	0.98	0.78	3.94 ± 0.08	4.25 ± 0.34

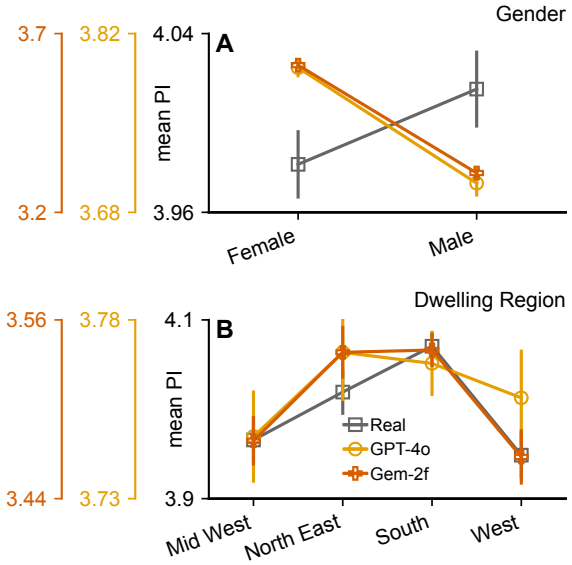

Figure 8: Mean purchase intent stratified by respondents' gender and dwelling region (shown are results from the SSR method for both GPT-4o and Gem-2f). Error bars represent standard errors.

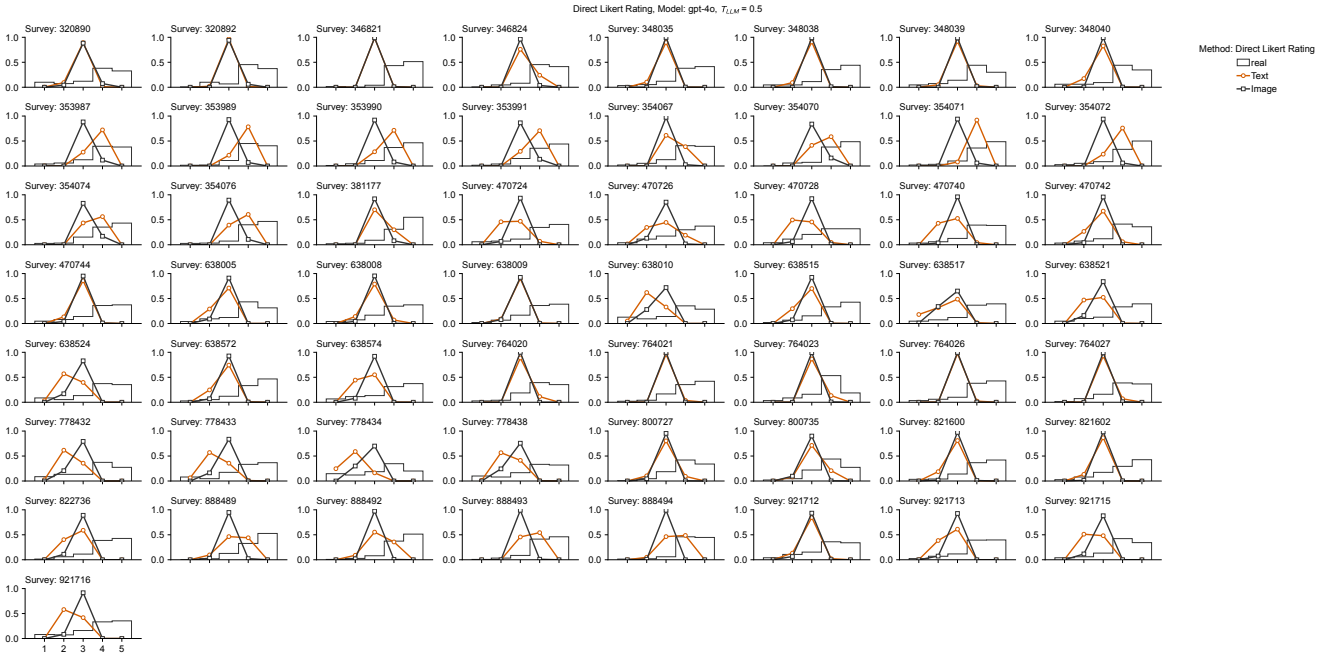

Figure 9: Survey histograms for direct Likert ratings at $T_{\text{LLM}} = 0.5$ for GPT-4o.

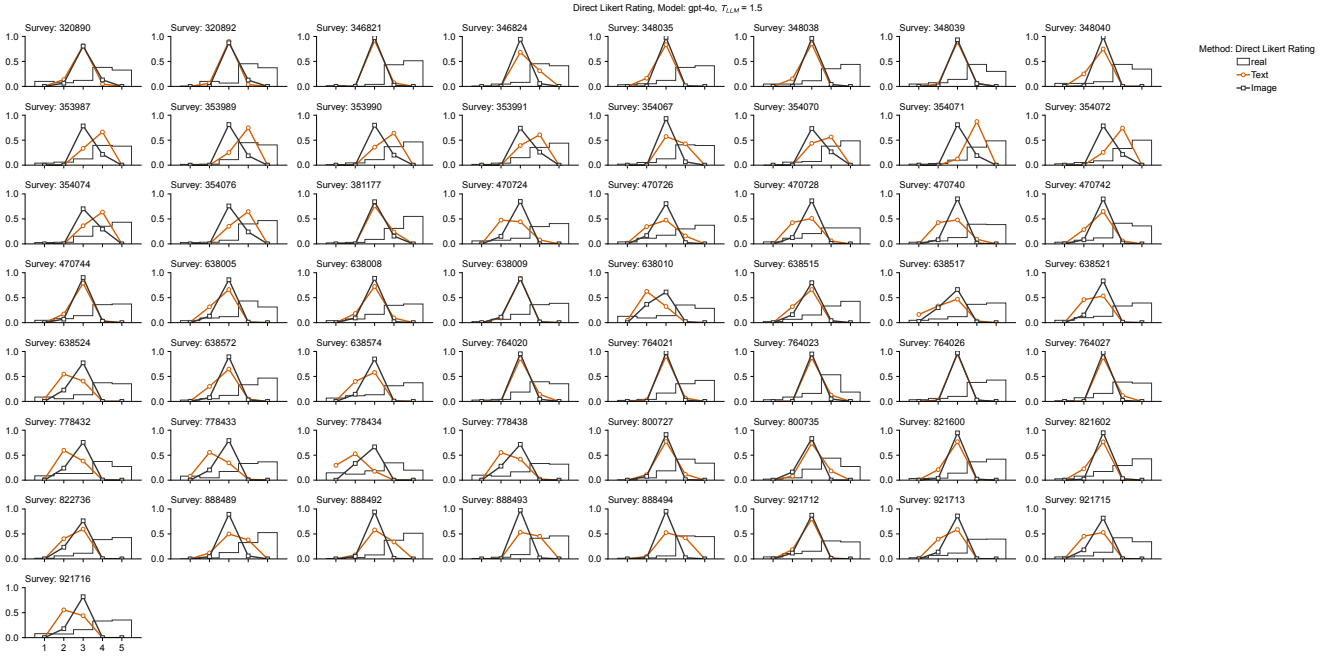

Figure 10: Survey histograms for direct Likert ratings at $T_{\text{LLM}} = 1.5$ for GPT-4o.

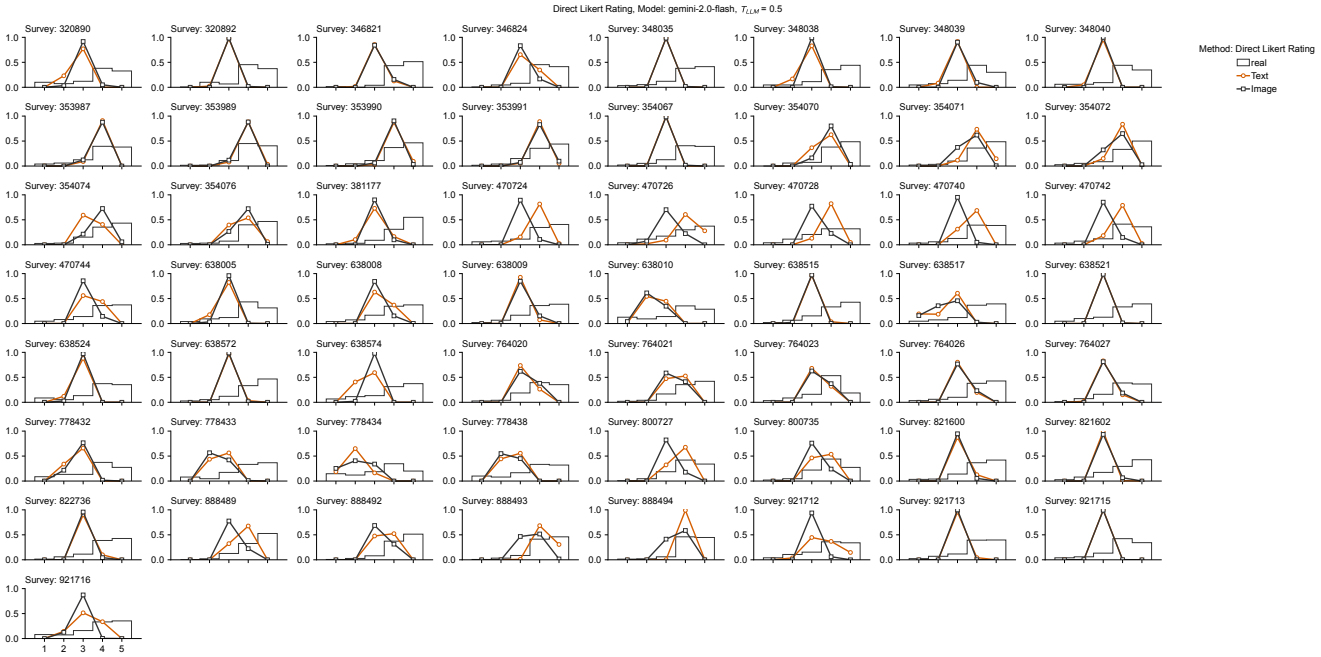

Figure 11: Survey histograms for direct Likert ratings at $T_{\text{LLM}} = 0.5$ for Gem-2f.

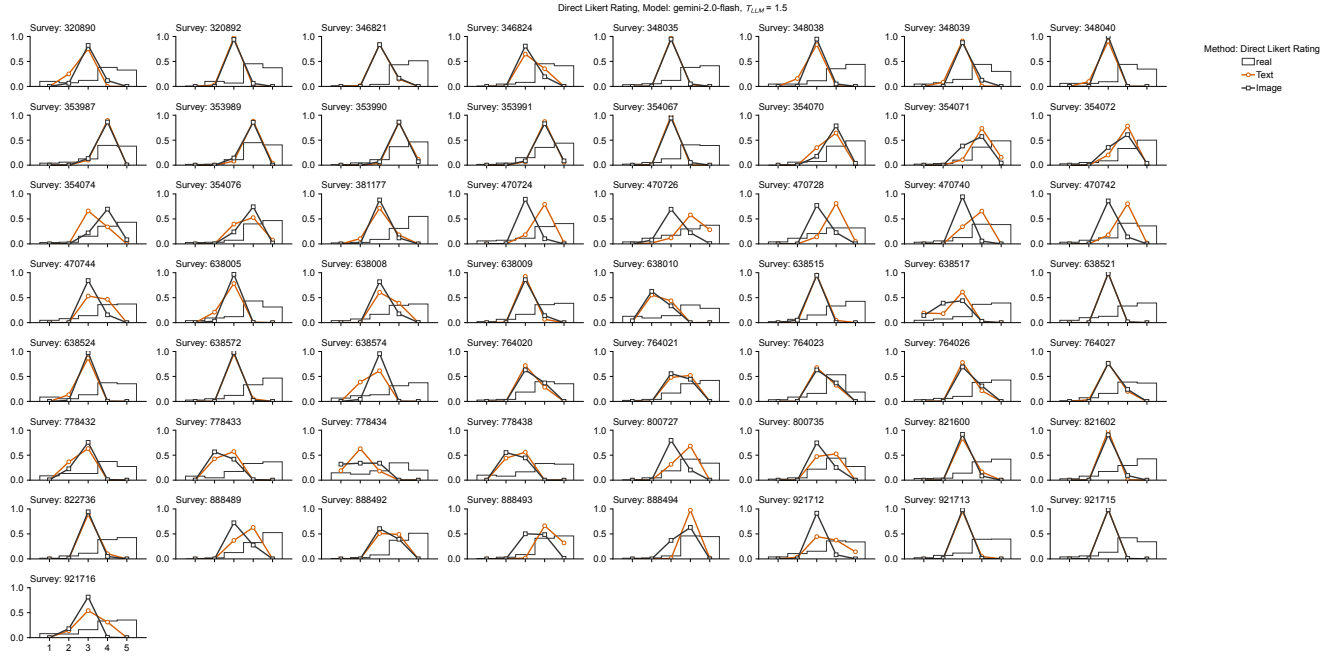

Figure 12: Survey histograms for direct Likert ratings at $T_{\text{LLM}} = 1.5$ for Gem-2f.

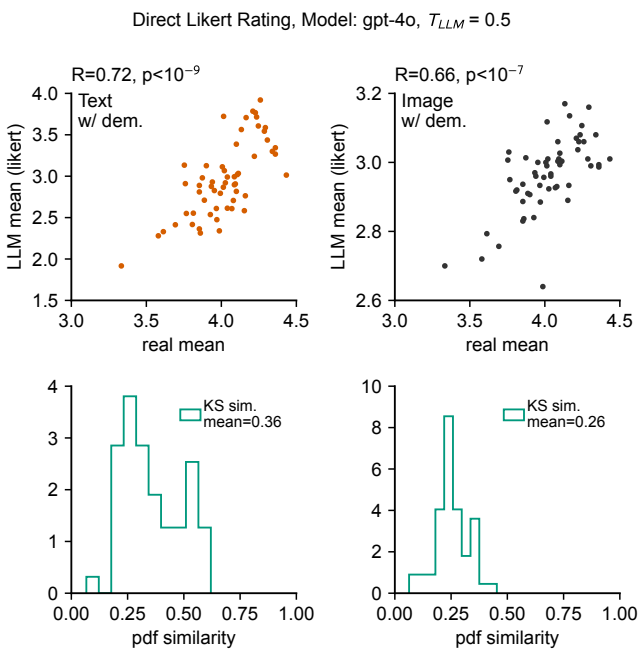

Figure 13: Success metrics for direct Likert ratings at $T_{\text{LLM}} = 0.5$ for GPT-4o.

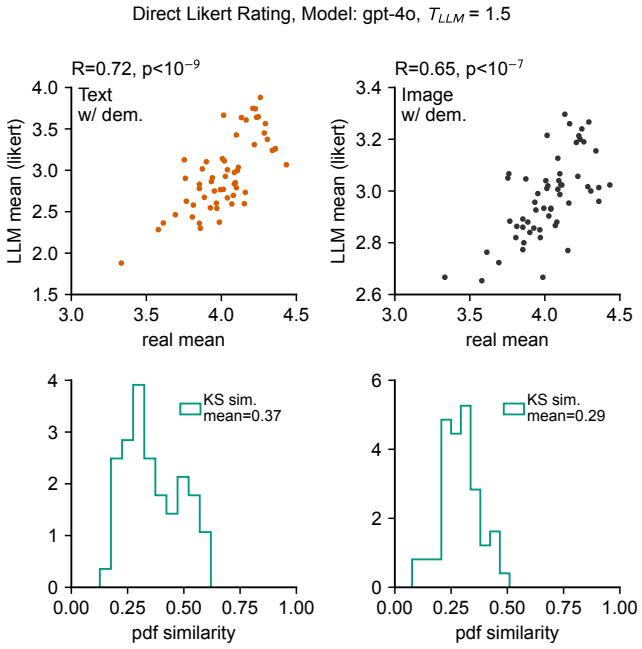

Figure 14: Success metrics for direct Likert ratings at $T_{\text{LLM}} = 1.5$ for GPT-4o.

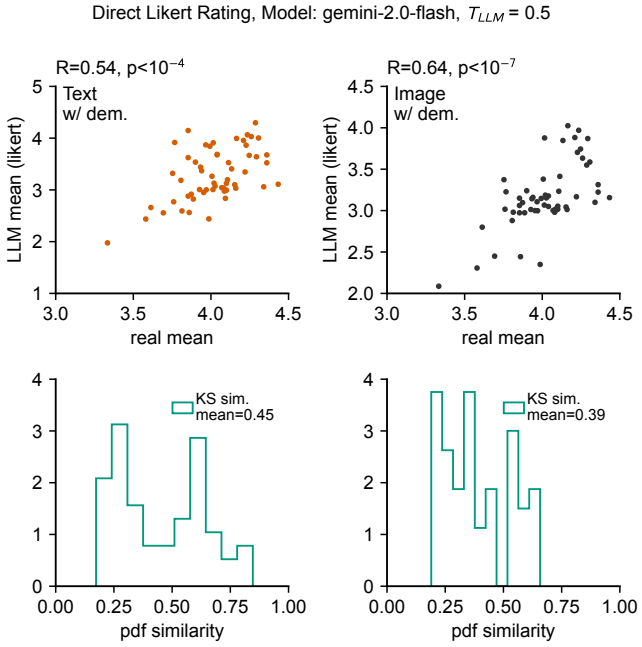

Figure 15: Success metrics for direct Likert ratings at $T_{\text{LLM}} = 0.5$ for Gem-2f.

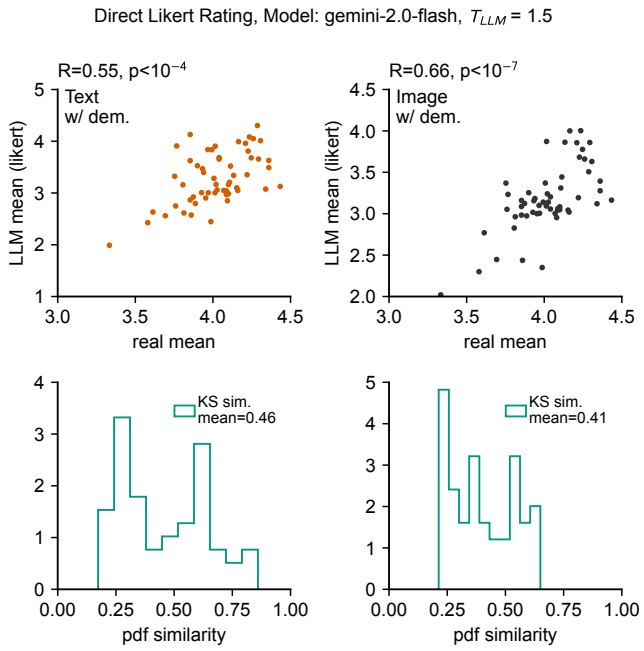

Figure 16: Success metrics for direct Likert ratings at $T_{\text{LLM}} = 1.5$ for Gem-2f.

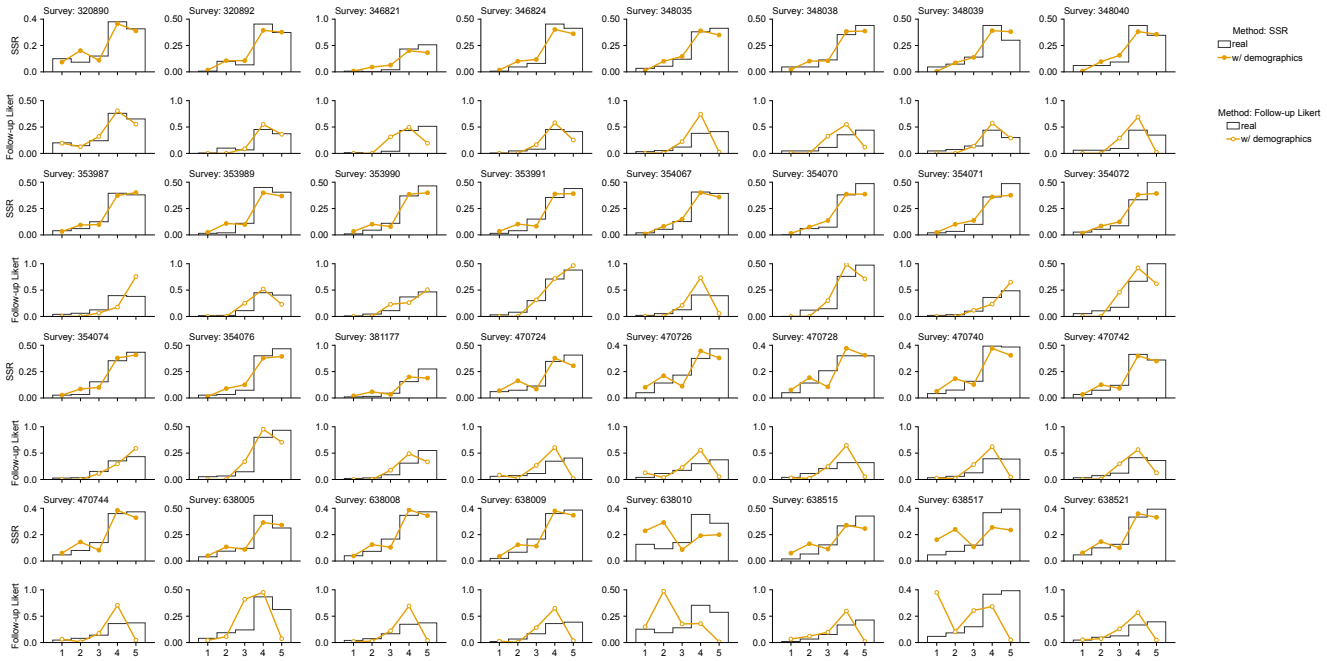

Figure 17: First set of survey histograms for textual elicitation with GPT-4o and follow-up ratings at $T_{\text{LLM}} = 0.5$, with image stimulus and full demography setup. For semantic similarity rating (SSR), we used the mean over all reference sets.

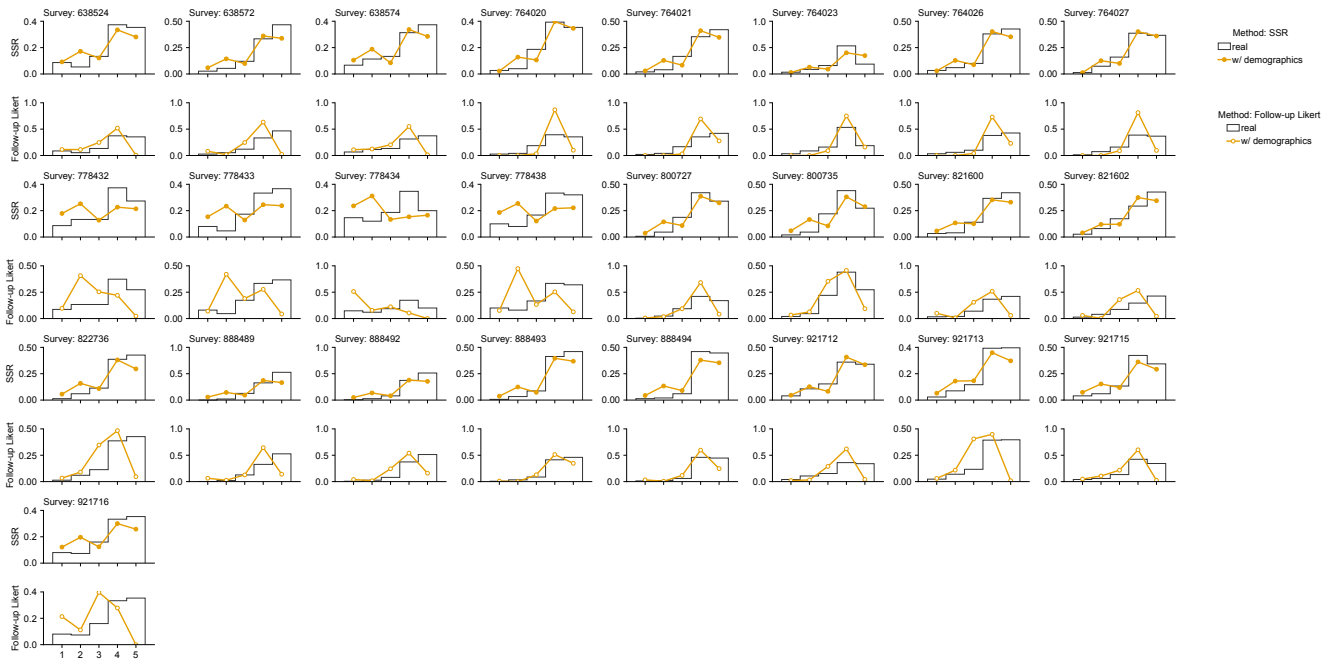

Figure 18: Second set of survey histograms for textual elicitation with GPT-4o and follow-up ratings at $T_{\text{LLM}} = 0.5$, with image stimulus and full demography setup. For semantic similarity rating (SSR), we used the mean over all reference sets.

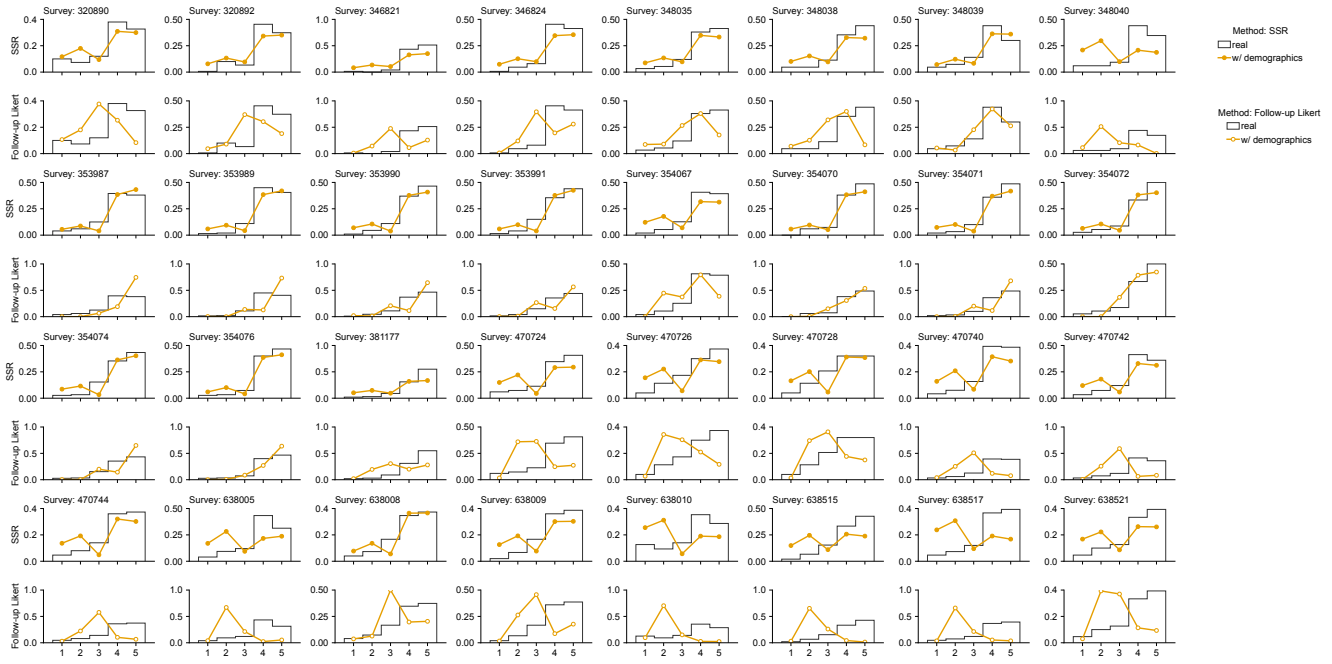

Figure 19: First set of survey histograms for textual elicitation with Gem-2f and follow-up ratings at $T_{\text{LLM}} = 0.5$, with image stimulus and full demography setup. For semantic similarity rating (SSR), we used the mean over all reference sets.

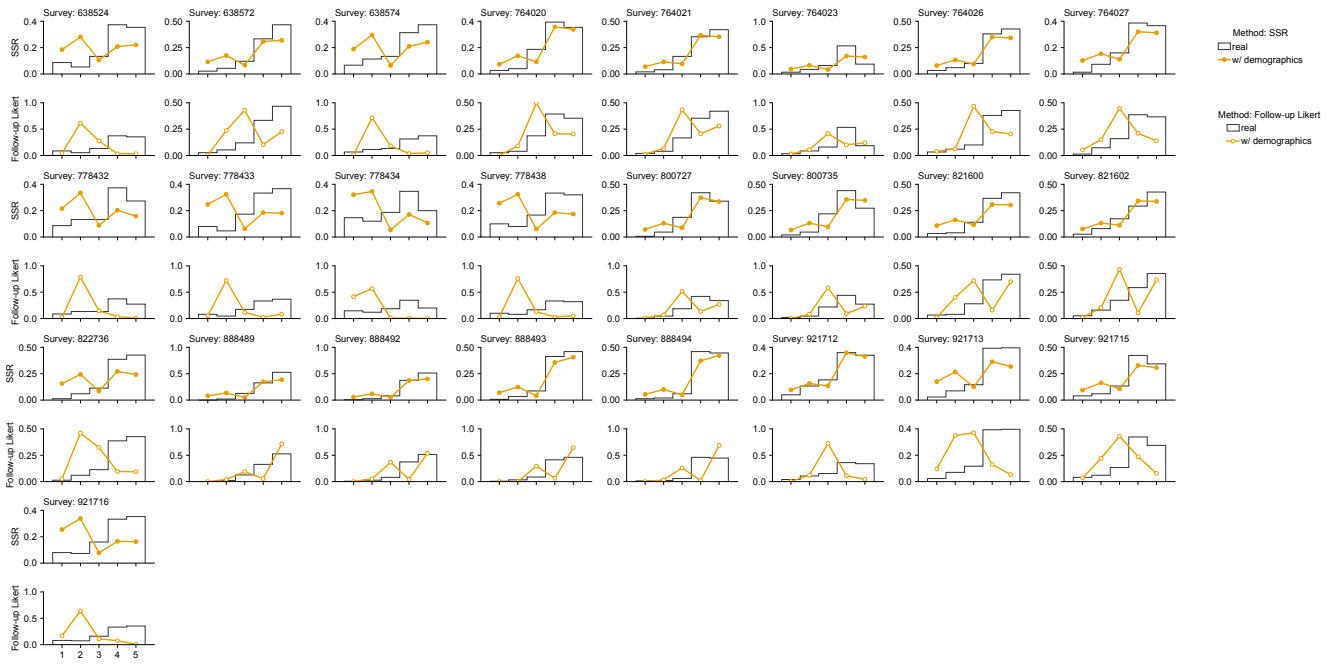

Figure 20: Second set of survey histograms for textual elicitation with Gem-2f and follow-up ratings at $T_{\text{LLM}} = 0.5$, with image stimulus and full demography setup. For semantic similarity rating (SSR), we used the mean over all reference sets.

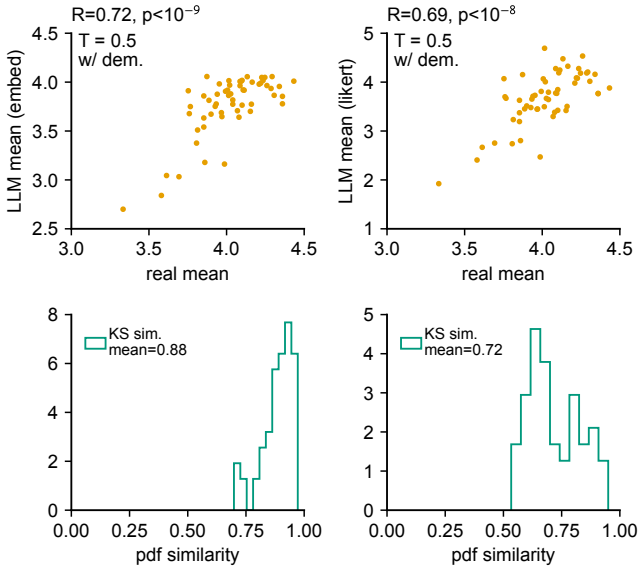

Figure 21: Success metrics for textual elicitation at $T_{\text{LLM}} = 0.5$ with GPT-4o, with image stimulus and full demography setup. For semantic similarity rating (SSR), we used the mean over all reference sets.

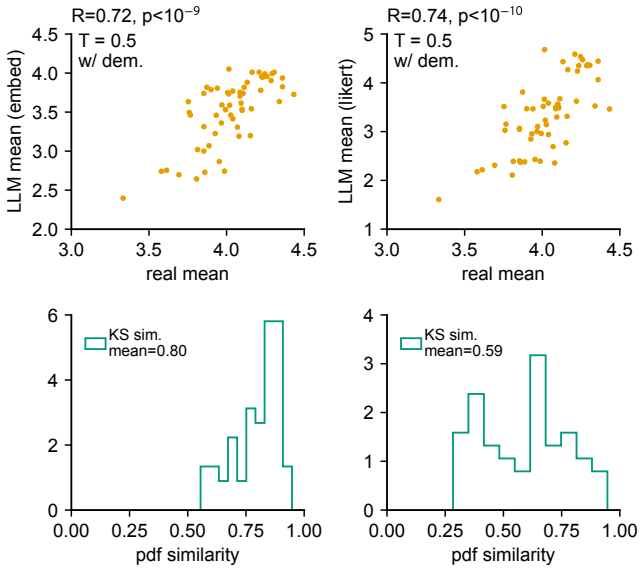

Figure 22: Success metrics for textual elicitation at $T_{\text{LLM}} = 0.5$ with Gem-2f, with image stimulus and full demography setup. For semantic similarity rating (SSR), we used the mean over all reference sets.

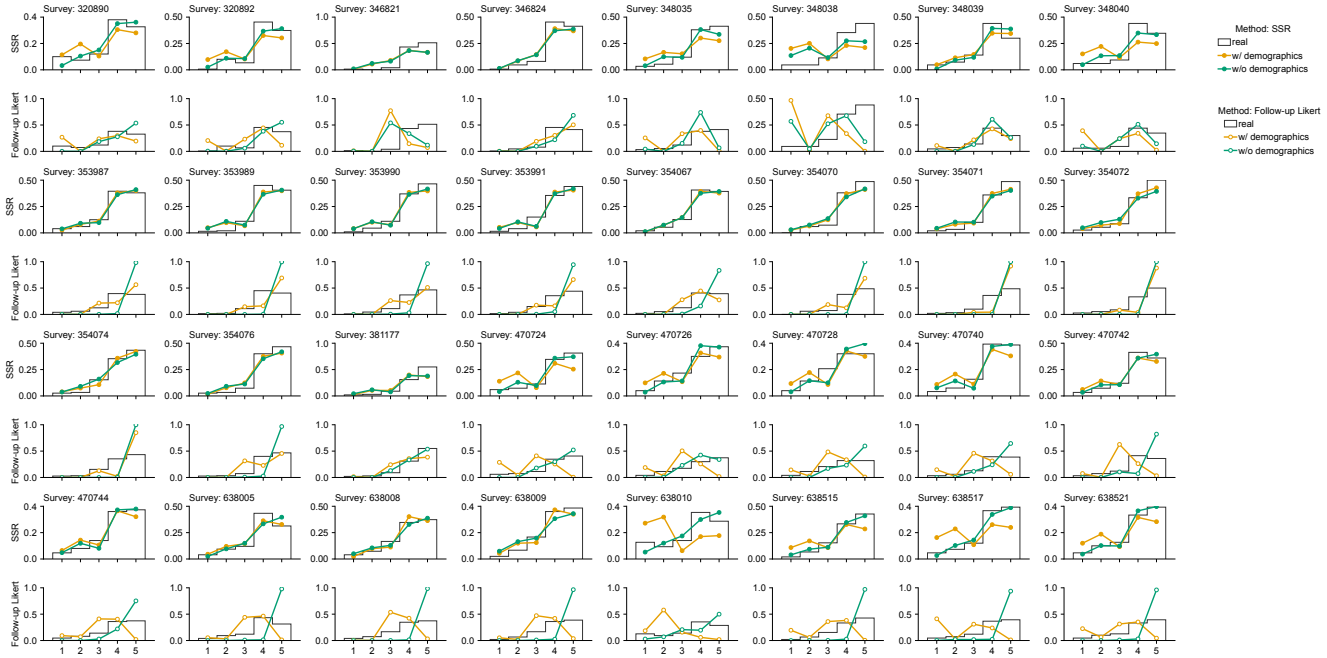

Figure 23: First set of survey histograms for textual elicitation with GPT-4o and follow-up ratings at $T_{\text{LLM}} = 0.5$, with text stimulus and alternating between prompting the LLM with *full* demographic information and *zero* demographic information. For semantic similarity rating (SSR), we used the mean over all reference sets.

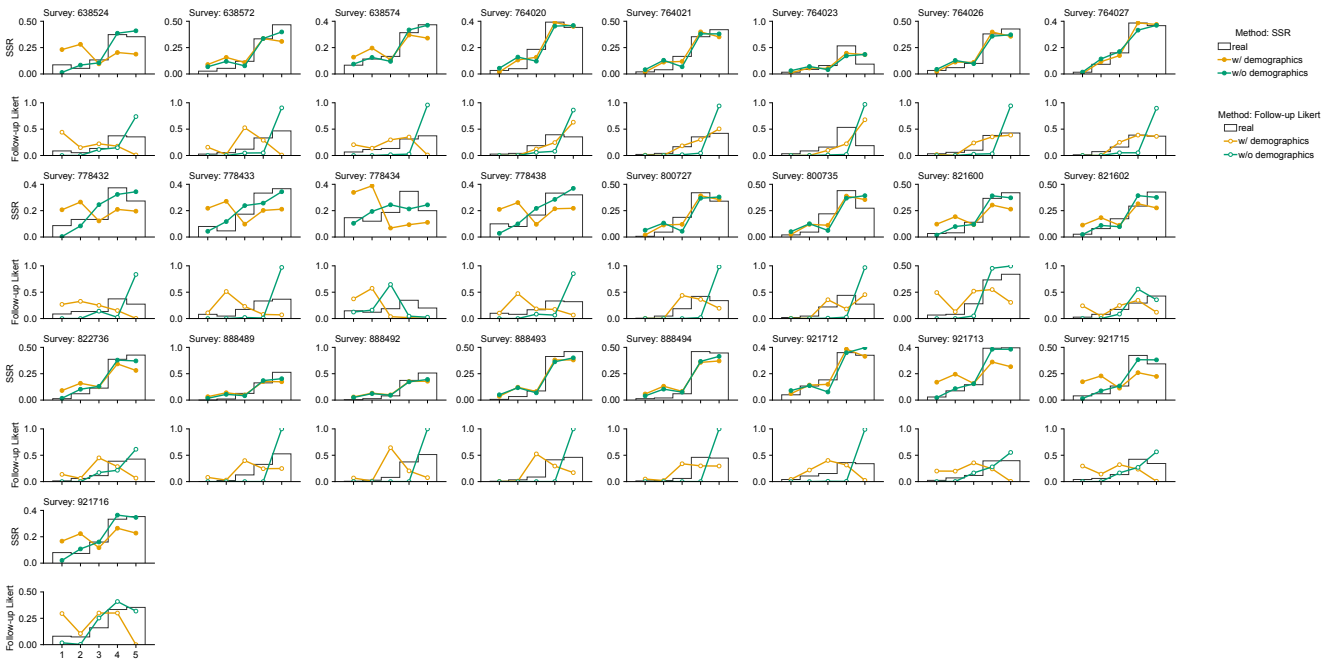

Figure 24: Second set of survey histograms for textual elicitation with GPT-4o and follow-up ratings at $T_{\text{LLM}} = 0.5$, with text stimulus and alternating between prompting the LLM with *full* demographic information and *zero* demographic information. For semantic similarity rating (SSR), we used the mean over all reference sets.

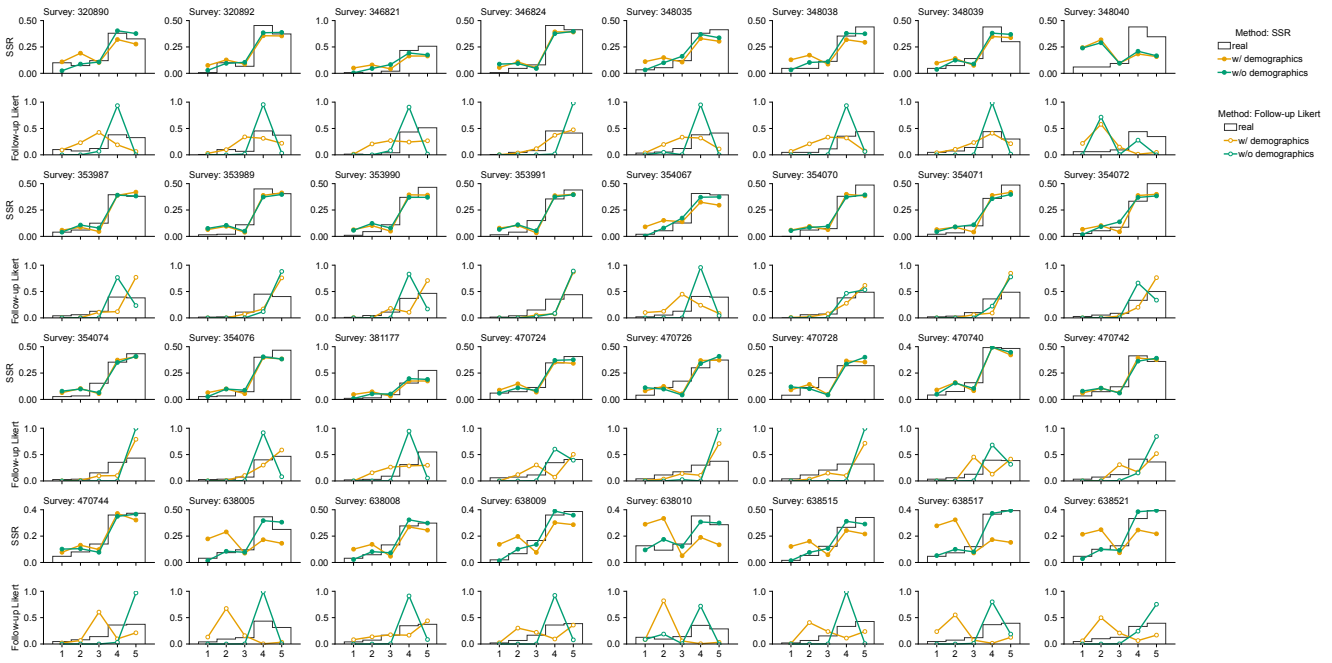

Figure 25: First set of survey histograms for textual elicitation with Gem-2f and follow-up ratings at $T_{\text{LLM}} = 0.5$, with text stimulus and alternating between prompting the LLM with *full* demographic information and *zero* demographic information. For semantic similarity rating (SSR), we used the mean over all reference sets.

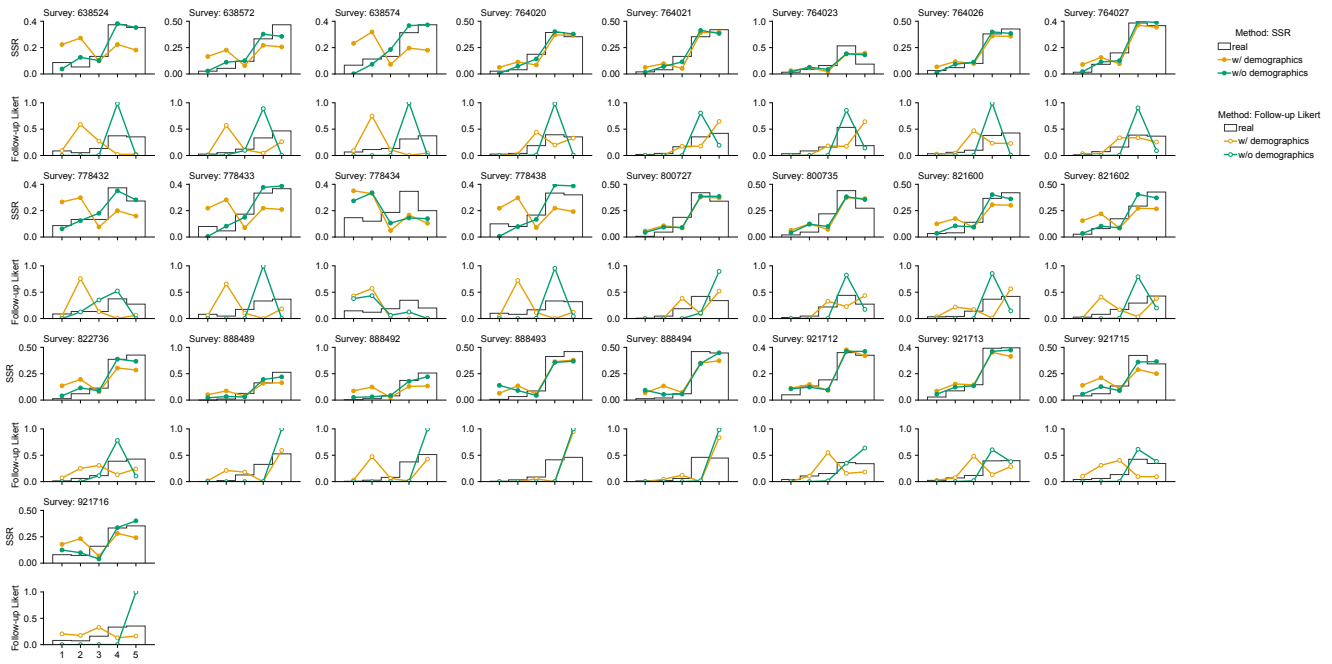

Figure 26: Second set of survey histograms for textual elicitation with Gem-2f and follow-up ratings at $T_{\text{LLM}} = 0.5$, with text stimulus and alternating between prompting the LLM with *full* demographic information and *zero* demographic information. For semantic similarity rating (SSR), we used the mean over all reference sets.

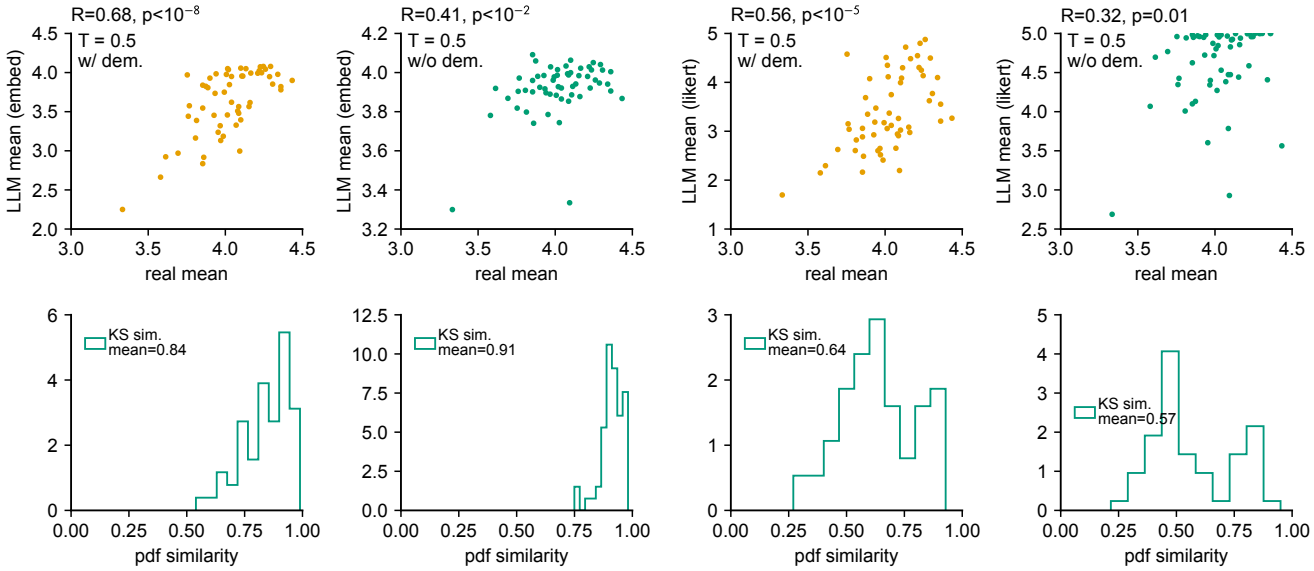

Figure 27: Success metrics for textual elicitation and demography experiments, at $T_{\text{LLM}} = 0.5$ with GPT-4o and with text stimulus. For semantic similarity rating (SSR), we used the mean over all reference sets.

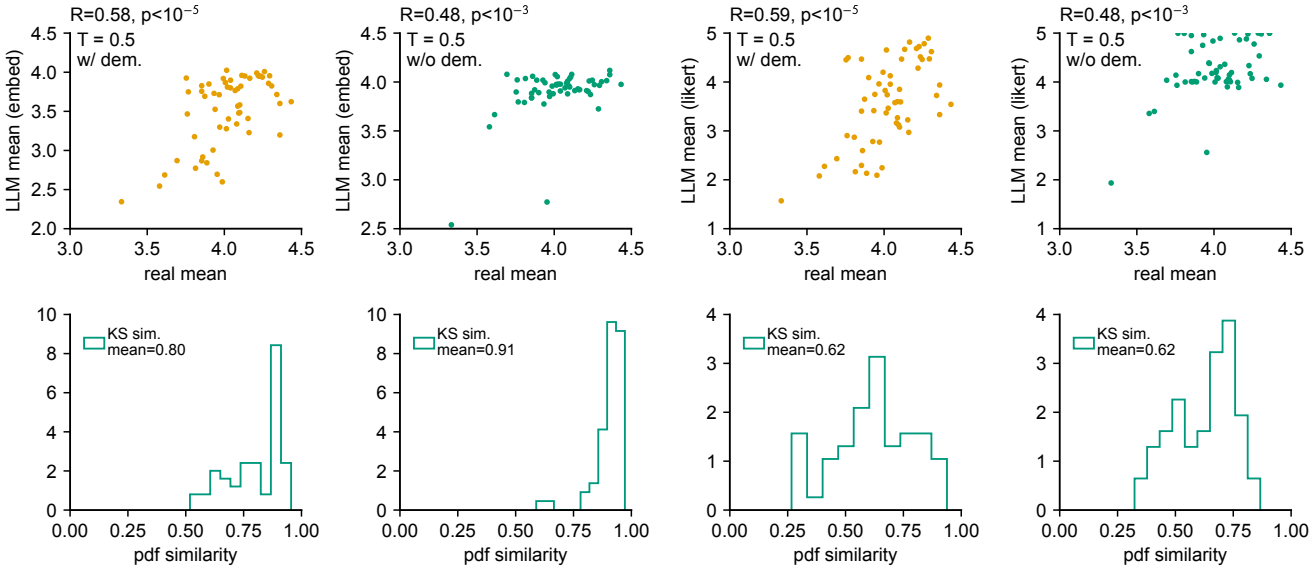

Figure 28: Success metrics for textual elicitation and demography experiments, at $T_{\text{LLM}} = 0.5$ with Gem-2f and with text stimulus. For semantic similarity rating (SSR), we used the mean over all reference sets.

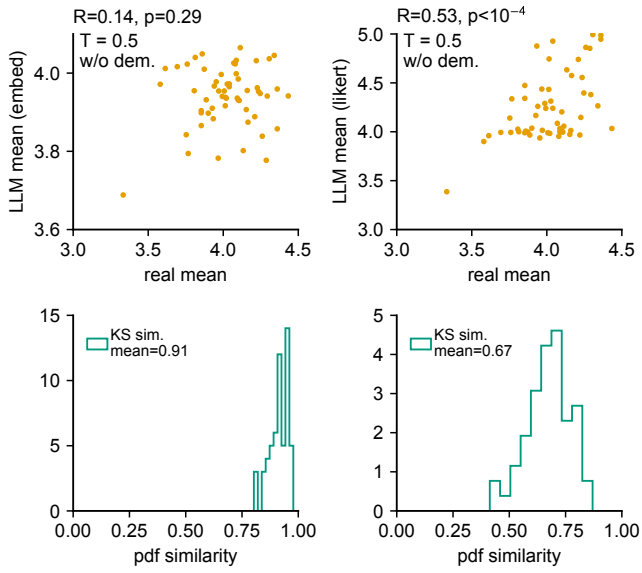

Figure 29: Success metrics for textual elicitation and demography experiments, at $T_{\text{LLM}} = 0.5$ with Gem-2f and image stimulus. For semantic similarity rating (SSR), we used the result for the best reference set (4).

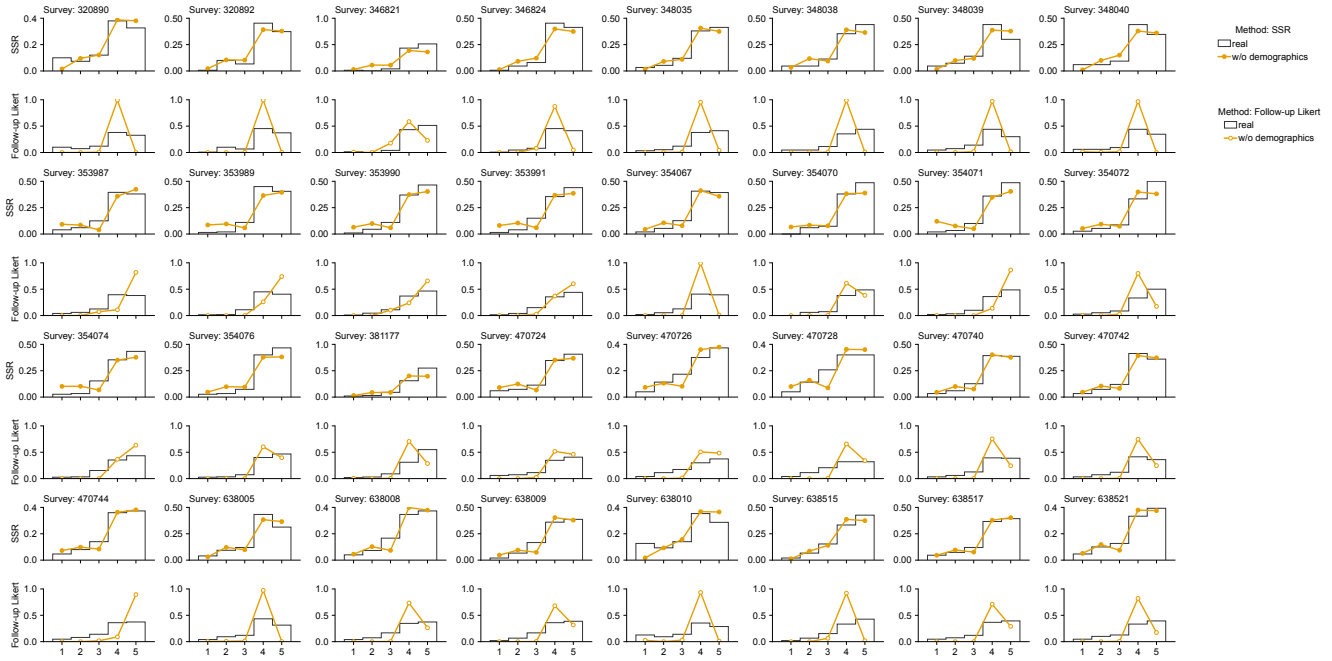

Figure 30: First set of survey histograms for textual elicitation with Gem-2f and follow-up ratings at $T_{\text{LLM}} = 0.5$, with image stimulus and prompting the LLM with *zero* demographic information. For semantic similarity rating (SSR), we used the the result for the best reference set (4).

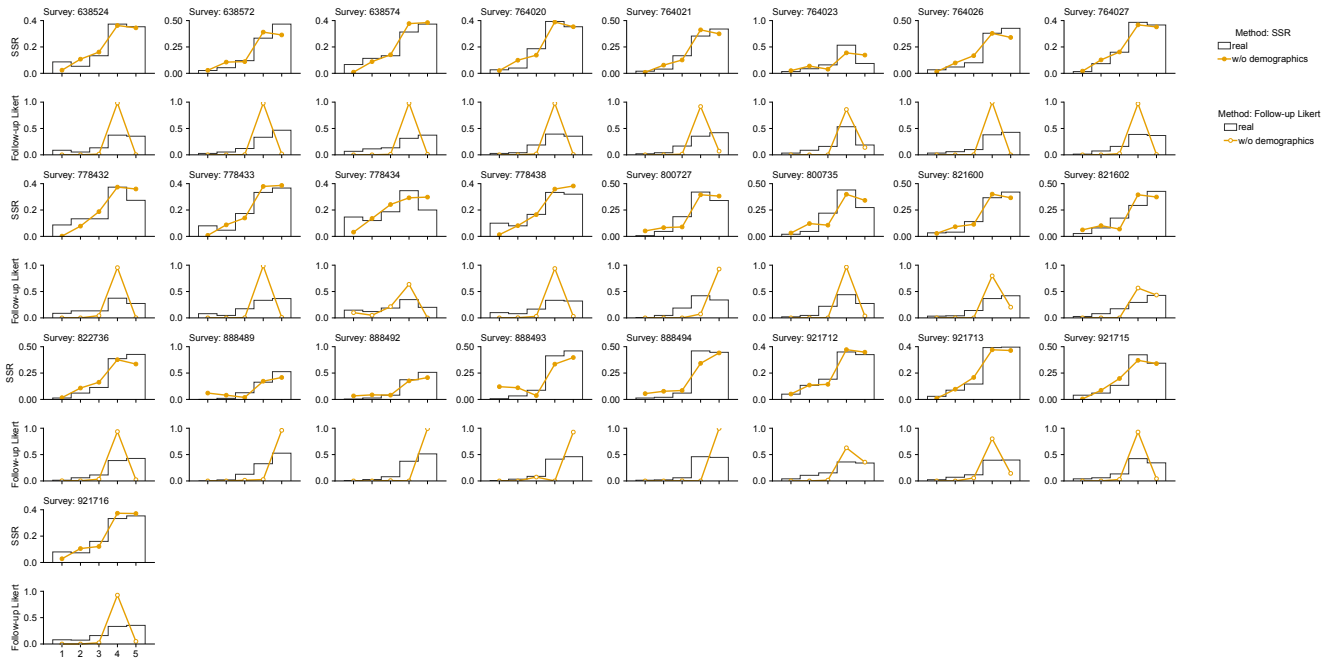

Figure 31: Second set of survey histograms for textual elicitation with Gem-2f and follow-up ratings at $T_{\text{LLM}} = 0.5$, with image stimulus and prompting the LLM with *zero* demographic information. For semantic similarity rating (SSR), we used the the result for the best reference set (4).

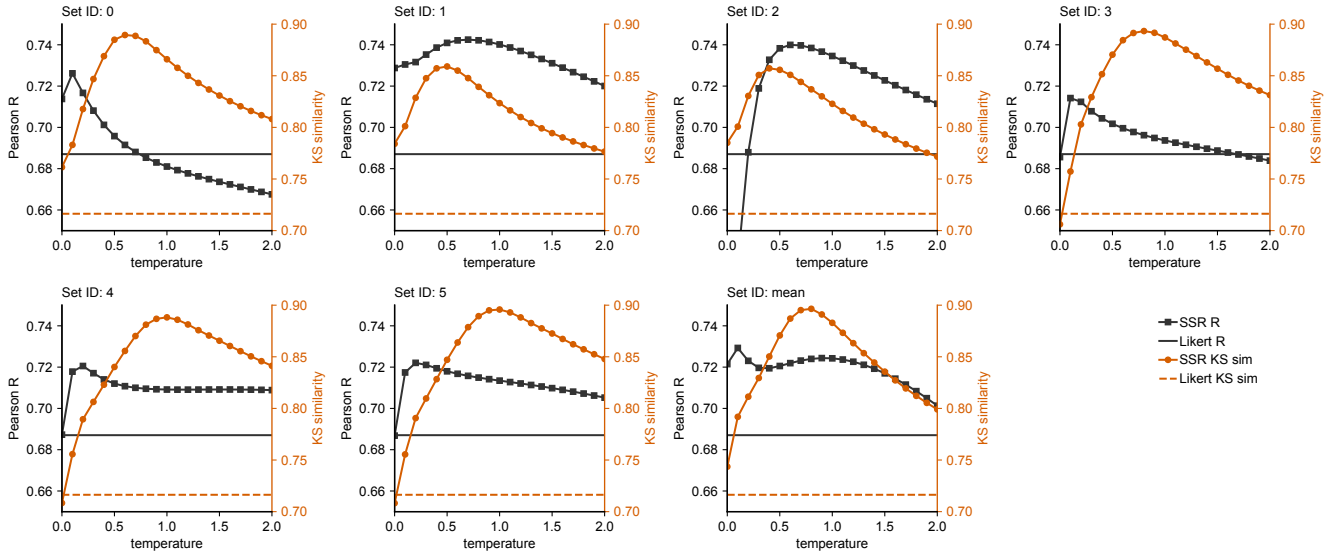

Figure 32: Scan over post-elicitation temperature T values and change in success metrics for textual elicitation at $T_{\text{LLM}} = 0.5$ with GPT-4o and image stimulus, with full demography setup. For semantic similarity rating (SSR), we used the mean over all reference sets. Horizontal lines refer to the corresponding follow-up Likert rating values for this experiment, which are unaffected by choice of reference set and T.

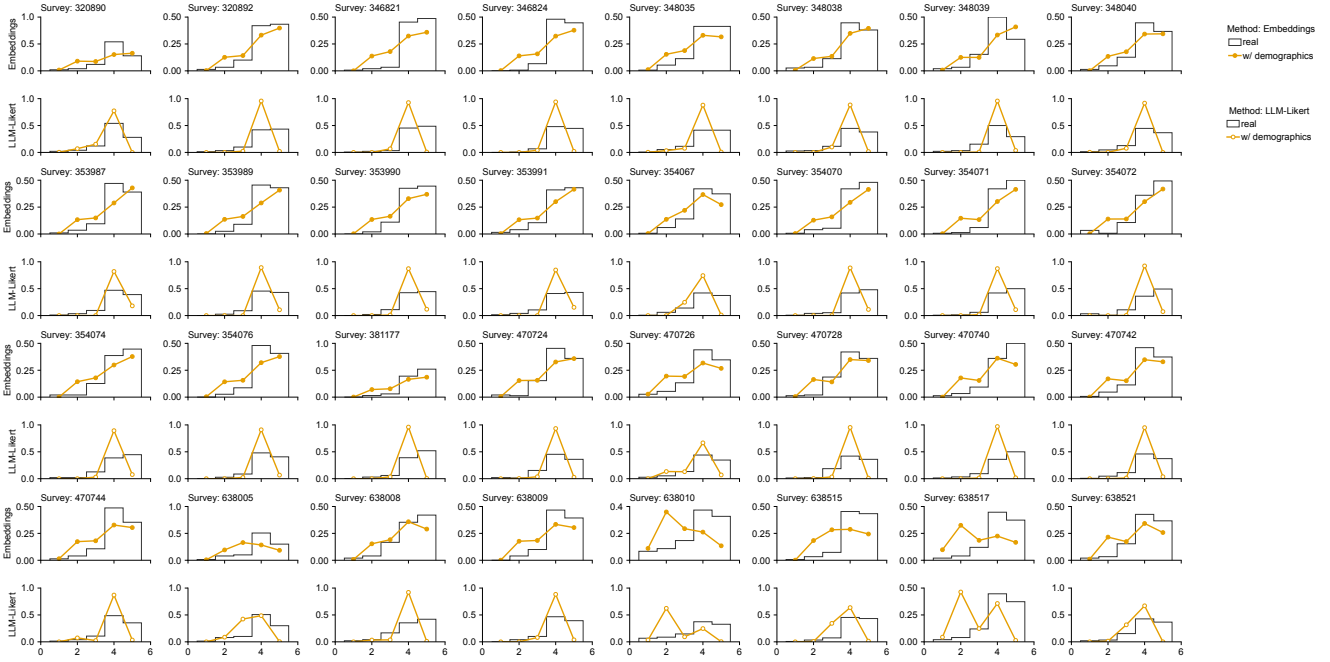

Figure 33: First set of survey histograms for textual elicitation to question "How relevant is this concept for you?" with Gem-2f and follow-up ratings at $T_{\text{LLM}} = 0.5$, with image stimulus and full demography setup. For semantic similarity rating (SSR), we used the mean over three reference sets that were constructed for this question specifically.

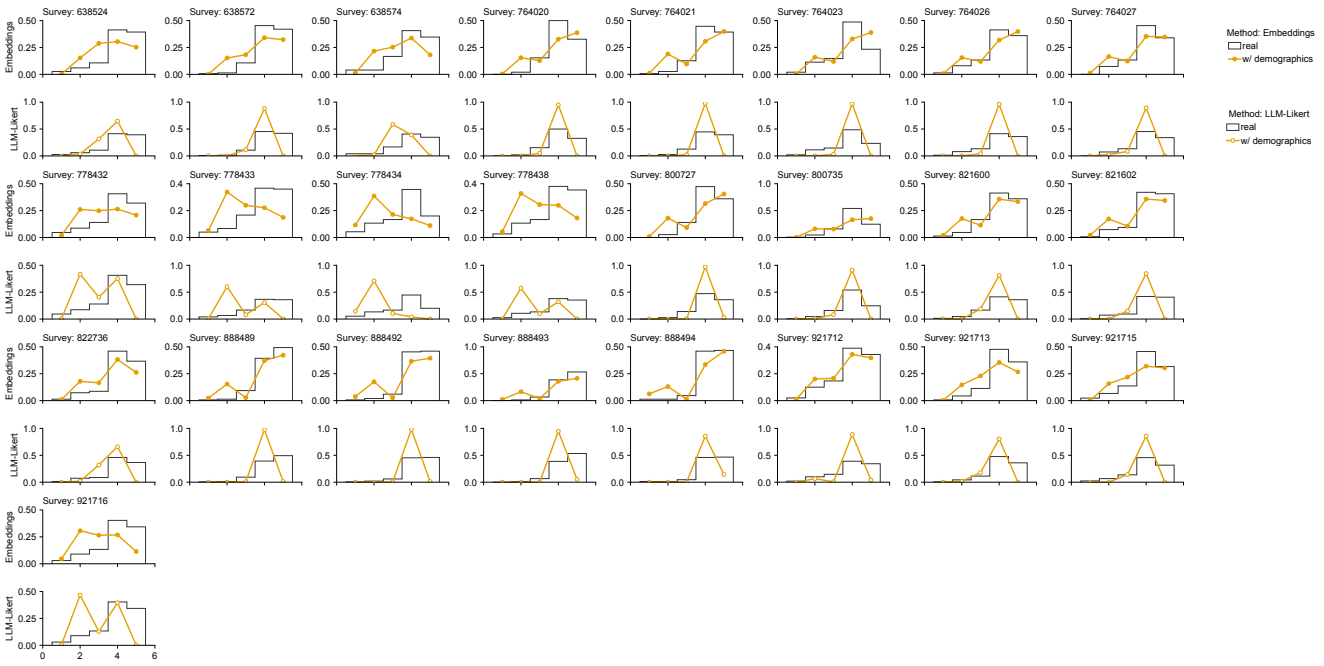

Figure 34: Second set of survey histograms for textual elicitation to question "How relevant is this concept for you?" with Gem-2f and follow-up ratings at $T_{\text{LLM}} = 0.5$, with image stimulus and full demography setup. For semantic similarity rating (SSR), we used the mean over three reference sets that were constructed for this question specifically.

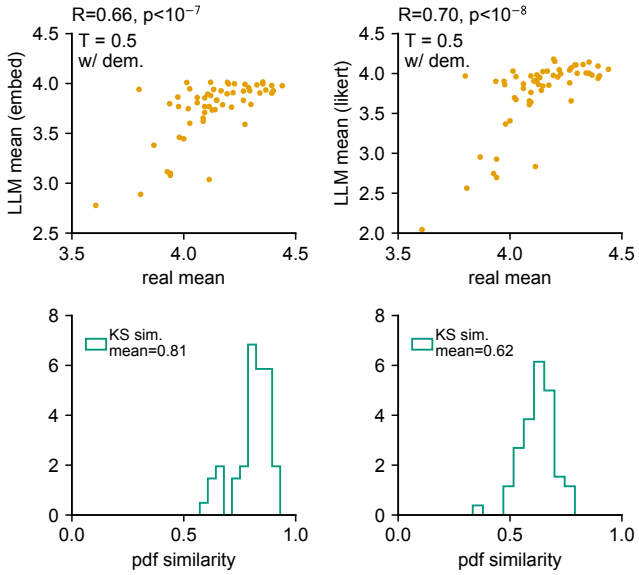

Figure 35: Success metrics for textual elicitation to question "How relevant is this concept for you?" with Gem-2f and follow-up ratings at $T_{\text{LLM}} = 0.5$, with image stimulus and full demography setup. For semantic similarity rating (SSR), we used the mean over three reference sets that were constructed for this question specifically.