Behavioral Research Through Interpretable, Dimensionality-reduced Generative AI Embeddings (BRIDGE): A Method to Incorporate Real-World Stimuli in Consumer Experiments

Anirban Mukherjee
Hannah Hanwen Chang
Sachin Gupta

April 20, 2026

Anirban Mukherjee (anirban@avyayamholdings.com) is Principal at Avyayam Holdings. Hannah H. Chang (hannahchang@smu.edu.sg; corresponding author) is Associate Professor of Marketing at the Lee Kong Chian School of Business, Singapore Management University. Sachin Gupta is the Henrietta Johnson Louis Professor of Marketing at the SC Johnson College of Business, Cornell University. This research was supported by the Ministry of Education (MOE), Singapore, under its Academic Research Fund (AcRF) Tier 2 Grant, No. MOE-T2EP40124-0005.

Abstract

Traditional experiments often rely on a few, stylized stimuli, which can limit realism and undermine generalizability beyond the sampled stimuli—known as the stimulus-sampling problem. To address these challenges, this paper introduces BRIDGE, a novel analytical method that enables the use of many unaltered real-world descriptions as experimental stimuli. Leveraging foundational generative AI embeddings, BRIDGE develops (1) structured, low-dimensional, and interpretable representations of focal constructs and (2) statistical controls for non-focal nuisance variations, facilitating causal inference. The authors validate the method through extensive Monte Carlo simulations, two coffee certification experiments, and a large-scale wine choice experiment (plus its validation study). The results show that BRIDGE recovers true parameters even when textual stimuli contain unobserved nuisance variations, and can effectively account for different sources of confounding. In a large-scale choice experiment, 1,000 participants evaluated approximately 50,000 unique product descriptions randomly sampled from a corpus of nearly 120,000. Results show that entirely incidental initial products can shape participants' subsequent preferences. By incorporating many unaltered product texts into experiments, BRIDGE enhances the realism, generalizability, and practical relevance of consumer research in information-rich environments. A detailed researchers' guide and Python package `bridge` are provided.

Keywords: Research Design, Interpretable Artificial Intelligence, Generative AI, Consumer Behavior, Marketing.

Real-world product descriptions can be complex, elaborate, and diverse. For instance, when choosing a vacation destination, Bali (Indonesia) might highlight its stunning beaches, ancient temples, and a diverse range of activities like surfing and yoga retreats. Nairobi (Kenya) may be portrayed through its lively city-meets-safari appeal, with bustling Maasai craft markets and the opportunity to spot lions, giraffes, and other wildlife in the nearby national park. Consumers may be presented with many such product descriptions in free-form, unstructured texts. That richness helps consumers anticipate experiences and evaluate products but poses a challenge for researchers: How much of this real-world complexity should experimental stimuli retain?

Using stimuli with rich detail can help activate the same psychological processes that affect real-world consumer choices, enhancing both experimental and mundane realism (Camerer 1997; Morales, Amir, and Lee 2017; Wilson, Aronson, and Carlsmith 2010).[1] However, the inherent variability of real-world descriptions can make it difficult to ensure comparability across stimuli. A category may have hundreds or thousands of product descriptions, which may differ in countless ways, from what is said ("content," e.g., attributes, claims, information) to how it is said ("form," e.g., words used, style, length). Causal inference requires isolating the effect of a focal element embedded in the unstructured text (e.g., surfing, cultural heritage, wildlife) while holding constant the many other features that co-vary with it.

Some researchers condense, abbreviate, and stylize real-world product descriptions. For example, previous studies presented vacation experiences to participants using stylized displays (e.g., "A = (average décor, $120 per night)"; Frederick, Lee, and Baskin (2014), Studies 1a–1s) or abbreviated descriptions (e.g., holiday destinations described by name only; Sharot, Velasquez, and Dolan (2010), Study 1). While this approach may exclude content integral to the original allure of the products, it allows researchers to isolate the causal influence of a focal aspect of product descriptions on consumer response by affording greater control over nuisance variables (Calder, Phillips, and Tybout 1981; Camerer 1997; Wilson, Aronson, and Carlsmith 2010). Yet this control

[1]Experimental realism is the degree to which the experiment engages participants psychologically; mundane realism is the extent to which a study mirrors real-life situations, tasks, and environments (Wilson, Aronson, and Carlsmith 2010).

comes at a cost: with only a handful of select stimuli per condition, the observed effects may be due to idiosyncrasies of the specific stimuli rather than the intended construct (Clark 1973; Pham 2013; Wells and Windschitl 1999). This is the stimulus-sampling problem, and variation in unstructured texts tend to amplify it: the potential stimulus space of unstructured texts is far larger than that of structured stimuli, making it difficult to sample representatively.

The stimulus-sampling problem undermines a study's internal validity (Simonsohn, Montealegre, and Evangelidis 2024), as the sample may be biased by uncontrolled confounds, inflating Type I error (Judd, Westfall, and Kenny 2012; Wickens and Keppel 1983). It threatens construct validity[2] (Wells and Windschitl 1999), as specific features of a sampled stimulus may introduce alternative construct interpretation, and external validity, as findings may not generalize beyond stimuli used (e.g., Baribault et al. 2018; Judd, Westfall, and Kenny 2012). Scholars have speculated that past failures to replicate experimental results may stem from the stimulus-sampling problem (e.g., Westfall, Judd, and Kenny 2015).

In this paper, we propose BRIDGE (Behavioral Research through Interpretable, Dimensionality-reduced Generative AI Embeddings), a novel data analysis methodology that enables a fundamentally different approach to stimulus sampling. Under this approach, participants in a single experiment can be exposed to distinct stimuli randomly selected from a corpus of real-world product descriptions, resembling a series of micro-experiments.

BRIDGE facilitates the analysis of participants' responses to diverse, unstructured stimuli while maintaining control across varying stimulus characteristics. It (1) uses a large language model (LLM) to generate high-dimensional, unstructured numerical embeddings of the unstructured stimuli and (2) transforms these embeddings into lower-dimensional, interpretable "knowledge representations" of the product attributes described in the stimuli, serving as numerical representations that are amenable to computational reasoning (Carvalho, Pereira, and Cardoso 2019; Levesque 1986; Tenenbaum et al. 2011). The lower dimensionality and interpretability of these

[2]As Wells and Windschitl (1999) point out, "the failure to sample stimuli can threaten construct validity…when 'the operations which are meant to represent a cause or effect can be construed in terms of more than one construct' (Cook and Campbell 1979, p. 59)" (p. 1116).

knowledge representations make them practical for use in theory-driven hypothesis testing with key, defined product attributes, as they can be incorporated in statistical models of participant behavior. BRIDGE also enables the development of statistical controls for nuisance variables to ensure stimulus comparability, extending conventional ANOVA/ANCOVA-style testing to settings with many varied real-world stimuli, while also encompassing conventional experimental designs as special cases. In doing so, it facilitates broader stimulus sampling, increases statistical power through multiple distinct product presentations per participant, and improves generalizability through greater coverage of product offerings in the market.

We organize our paper as follows. We develop BRIDGE, describing the data structures it addresses. We present a synthetic-data study identifying key boundary conditions. We demonstrate our methodology using multiple experiments. The first two experiments use controlled stimuli and conventional designs to provide evidence that the inference obtained from BRIDGE (without explicit confound specification) is comparable to an "unconfounded" benchmark that uses direct knowledge of the confound structure to yield consistent and efficient estimates. In the third experiment, we illustrate the use of BRIDGE in addressing a novel consumer research question that would be challenging to investigate using conventional experimental designs. This laboratory experiment involves 1,000 participants who are each presented with the real-world descriptions of 32 pairs of wines randomly sampled from nearly 120,000 available on the market; each wine description averaging 53 words (SD = 11.86) and the experiment encompassing almost 50,000 unique wines. The inclusion of such rich and diverse stimuli would be impracticable using traditional experimental designs with the same number of participants. We present validation studies comprising both perturbation analyses and follow-up experiments to further support our findings. We conclude with a discussion of the potential applications of BRIDGE methodology, its limitations, and possible directions for future research.

BRIDGE: METHOD DEVELOPMENT

The traditional experimental approach—using a few, simplified, and stylized stimuli—is subject to the stimulus-sampling problem. Identified as the "language-as-a-fixed-effect fallacy" (Clark 1973) and later framed as the stimulus-sampling problem (Wells and Windschitl 1999), this issue arises when idiosyncratic features of the selected stimuli are confounded with the intended construct. It threatens a study's *internal validity* when the selected stimuli produce effects for reasons other than the hypothesized one, such as through uncontrolled confounds (Simonsohn, Montealegre, and Evangelidis 2024). It undermines *construct validity* when the specific features of the stimuli offer alternative construct explanations for an effect, making it unclear whether the operationalized variable accurately reflects the intended theoretical construct (Cook and Campbell 1979; Wells and Windschitl 1999). It challenges *external validity*, as the findings may not generalize to the larger population of stimuli in the category (Baribault et al. 2018; Pham 2013; Westfall, Judd, and Kenny 2015). In addition to these validity concerns, treating stimuli as fixed rather than random when stimuli are sampled from a broader population inflates Type I error rates (Judd, Westfall, and Kenny 2012). Furthermore, these issues are amplified by the file drawer problem (Rosenthal 1979): if stimuli that yield significant results are more likely to be reported, they are disproportionately likely to be adopted in future studies. This selective reuse truncates the distribution from which stimuli are drawn, further inflating the probability of false conclusions and skewed inference. Past failures to replicate experimental results have been speculated to stem, in part, from the stimulus-sampling problem (e.g., Westfall, Judd, and Kenny 2015).

Recent work has proposed managing the stimulus-sampling problem *ex ante* at the design stage. The "Mix-and-Match" framework (Simonsohn, Montealegre, and Evangelidis 2024) provides a systematic procedure for stratified sampling of diverse stimuli within each condition and matching them across conditions to manage confounds. Its accompanying "Stimulus Plots" enable exploratory assessment of stimulus variation by comparing observed heterogeneity in results to that expected under homogeneity.

Several features of the questions and contexts that BRIDGE targets, however, present challenges for such design-centered approaches. When stimuli are unstructured real-world descriptions, the nuisance dimensions (non-focal variation) along which they vary can be numerous, latent, and difficult to identify *a priori*, let alone measure and balance across conditions. In addition, when stimuli vary simultaneously on multiple dimensions (e.g., attributes), the number of potential strata grows multiplicatively with each added dimension. While Simonsohn, Montealegre, and Evangelidis (2024) recommend a small number of categories (e.g., five) rather than exhaustive crossing, selecting which dimensions to stratify on requires knowing which dimensions matter, which may be difficult to obtain with unstructured stimuli. Stimulus Plots provide a valuable diagnostic for *detecting* stimulus variation, but their diagnostic value also depends on having sufficient observations per stimulus, a requirement that becomes harder to meet as the number of stimuli in a study grows.

BRIDGE complements design-based approaches by addressing these challenges *ex post* at the analysis stage, in contexts where researchers aim to test hypotheses using many unstructured textual stimuli (e.g., real-world wine tasting notes, financial product summaries, and sustainability initiative narratives). BRIDGE offers a unified model that pools information across stimuli while accounting for non-focal variation. Rather than requiring the careful curation of select stimuli before an experiment, it enables researchers to sample at scale from large, real-world corpora. It does so by using an interpretable AI model to decompose each stimulus into (1) structured, low-dimensional representations of the focal attributes and (2) statistical controls for latent nuisance variables.

In short, where Mix-and-Match manages confounds by carefully curating *what* is shown, BRIDGE manages them by algorithmically controlling for variation in the analysis of *what was* shown. It scales to settings where curation in advance may be difficult, such as when stimuli are many distinct real-world product descriptions. To understand BRIDGE, consider the standard analytical framework that it builds upon. In a traditional experiment, responses are analyzed

using a linear model:

$$y_i = \alpha + \beta D_c(i) + \gamma \text{covariates}(i) + \delta \text{attributes}(i) + \epsilon_i, \tag{1}$$

where y_i represents the response of participant i; α is the intercept; β is the hypothesized effect; γ are coefficients for covariates describing systematic factors; and δ are coefficients for stimulus attributes. $D_c(i)$ is a dummy variable indicating whether participant i is in condition c, covariates(i) are participant covariates (e.g., participant characteristics), and attributes(i) are attributes of interest (e.g., the emotional tone in which a description is written); the latter two pre-defined, based on theory. For simplicity, we omit a task index t, though this framework can accommodate studies with multiple tasks, task-varying effects, and task-specific stimuli (i.e., task-varying β and γ), in addition to accommodating interaction terms between the condition dummies and the covariates of interest.

In many experiments covariates$(i) = 0$ and attributes$(i) = 0$, and hypothesis testing on β is equivalent to ANOVA. If covariates$(i) \neq 0$, hypothesis testing on β is equivalent to ANCOVA, which assesses if the group means differ while accounting for the covariates. If attributes$(i) \neq 0$, it is typical to employ a regression equation and use Wald tests to examine if β, γ, or $\delta = 0$. This reduces to a t-test on β when considering only the treatment effect.

The stimulus-sampling problem arises directly from a key assumption in this approach: any systematic difference in responses must be attributable solely to the experimental manipulation $D_c(i)$ or the covariates(i). When participants are presented with diverse, unstructured real-world stimuli, the stimuli inevitably vary in numerous ways beyond the focal attributes. Not accounting for these extraneous variations (known as nuisance variables) violates that assumption, confounding measurement by offering alternative explanations.

Consequently, experimenters often choose few, simplified, and/or stylized stimuli. This approach improves comparability across conditions, with idiosyncratic (non-systematic) variation absorbed by the error term, ϵ_i. However, in prioritizing experimental control to protect internal

validity, this approach often simplifies stimuli and variety, diminishing construct and external validity.

Paradoxically, this approach also undermines the internal validity that it seeks to protect. Failing to account for the random variability introduced by the stimulus sample leads to underestimated error terms and, consequently, inflated "false positive" rates—increasing the likelihood of concluding an effect exists when it does not (Judd, Westfall, and Kenny 2012; Westfall, Judd, and Kenny 2015; Wickens and Keppel 1983).

One strategy to address these issues is to introduce many real-world descriptions as stimuli and manually code the nuisance variables, translating them into control variables. However, this strategy faces several difficulties. First, the relevant nuisance variables may not be known *a priori*—the many nuanced dimensions along which real-world descriptions differ (e.g., tone, style, voice, formality) are often hard to identify and isolate (Clark 1973). For instance, one product description might use vivid and emotive language to create an immersive experience, while another might rely on technical jargon or minimalist phrasing to convey sophistication. Such characteristics may be irrelevant to the research question yet still influence participant responses. Second, even if identifiable, the number of nuisance dimensions can be large, and including a control for each can quickly exhaust the information in the data. Consider, for example, the number of possible brands in the market for a typical consumer product such as running shoes; controlling for brand may require many fixed effects in a conventional model. Third, the sheer scale of coding may be prohibitive. Our wine study, for example, used 50,000 distinct stimuli selected from a set of 120,000 descriptions—a volume that would be difficult via manual coding.

BRIDGE employs an alternative strategy: it uses AI to systematically extract knowledge representations for key attributes and statistical controls for the nuisance variables and incorporate them into the analysis. This enables the specification of models of the form:

$$y_i = \alpha + \beta D_c(i) + \gamma \text{covariates}(i) + \delta \text{attribute}(i) + \zeta \text{controls}(i) + \epsilon_i, \tag{2}$$

where ζ represents coefficients for the statistical controls of nuisance variables, and controls(i) are the controls, which are algorithmically extracted from the stimuli. controls(i) are distinct from attributes(i) in that, whereas attributes(i) capture observed focal differences of interest that are often manipulated or controlled in traditional experiments, controls(i) represent variables that are algorithmically extracted to absorb the latent, unobserved nuisance variation.

To illustrate, consider an experiment where the stimuli vary in color. If color were known, a researcher could control for it by including it in the model, thereby ruling it out as a confounding factor when estimating β. Now, suppose color were unknown *a priori*. A strategy might be to use an algorithm to automatically code color, thereby enabling a similar analytical approach. This is analogous to BRIDGE but differs in scale and complexity: rather than a single observable attribute like color, BRIDGE is designed to identify and control for multiple, latent nuisance dimensions—tone, style, formality, and more—simultaneously.

If stimuli of many colors are introduced, adding a coefficient for each color leads to the inclusion of many coefficients. A fixed-length numerical representation (for color, RGB—a three-dimensional representation) leads to a more efficient econometric model. As BRIDGE is designed for real-world descriptions that can vary in many ways, it similarly develops low-dimensional representations of the potentially high-dimensional nuisance variables in stimuli as statistical controls.

Section Roadmap. Next, we provide a detailed explanation of how BRIDGE operates. We outline the general step-by-step procedure for conducting a study using BRIDGE and introduce perturbation analysis as a technique for validating the methodology. To facilitate adoption, we provide a web appendix with a detailed researcher's guide, introducing a Python package that includes all analytical components (see Web Appendix §B). Last, we discuss how BRIDGE's structured, algorithmic approach enhances research objectivity and reproducibility, facilitating research practices like pre-registration.

BRIDGE: An Interpretable AI Model for Theory Testing

BRIDGE is developed to address two challenges in the use of semantic embeddings (large language models [LLM]) for hypothesis testing. First, semantic embeddings are inherently unstructured—they lack explicit organization. Consequently, they lack interpretability: individual dimensions of the encoding do not correspond to specific, understandable features, making it challenging to relate these dimensions to theoretical constructs of interest or to link them directly to participant responses. Second, they are high-dimensional, often consisting of thousands of dimensions, which can pose computational challenges and reduce statistical power—an issue that is particularly crucial when working with typical sample sizes in laboratory studies.

Given a set of theory-defined focal attributes, BRIDGE uses a fully connected feedforward network to compress a high-dimensional, unstructured semantic embedding into a lower-dimensional intermediate representation (Johnson 1984).[3] As illustrated in Figure 1, the compression network consists of two stages: a projection layer and a reduction layer. The projection layer initially reduces the dimensionality of the embedding by exploiting the structure of the encoding; in the case of Matryoshka representations (Kusupati et al. 2022), this is simply a masking layer, as the embeddings have a nesting structure that permits truncation without information loss; in other architectures, a learnable projection to initially reduce dimensionality may be useful. The reduction layer, a feedforward network whose depth and width are discoverable through automatic tuning [e.g., Optuna; Akiba et al. (2019)], further compresses the projected embedding to produce a low-dimensional intermediate representation. A partitioned feedforward network maps this intermediate representation to an interpretable layer that is partitioned into distinct components, each of which is dedicated to predicting a distinct attribute during training.[4] The network learns to further compress the intermediate representation and extract attribute-relevant information

[3]BRIDGE is embedding-agnostic: any semantic embedding can serve as input. However, the quality of the input embeddings affects the precision of the extracted representations. For instance, using BERT embeddings (768 dimensions) in place of OpenAI text-embedding-3-large (3,072 dimensions) yields anchoring coefficients in the expected direction but attenuated and nonsignificant ($\hat{\beta} = 0.011$–0.015 vs. 0.038–0.039). See Web Appendix §F.6 for details.

[4]This approach is different from (standard) feedforward networks where the complete output of a preceding layer is used to predict an attribute (Caruana 1997).

(Bishop 1995; Hornik, Stinchcombe, and White 1989), embedding it in an attribute-specific vector space. The output from each component of the interpretable layer serves as the corresponding attribute's representation.

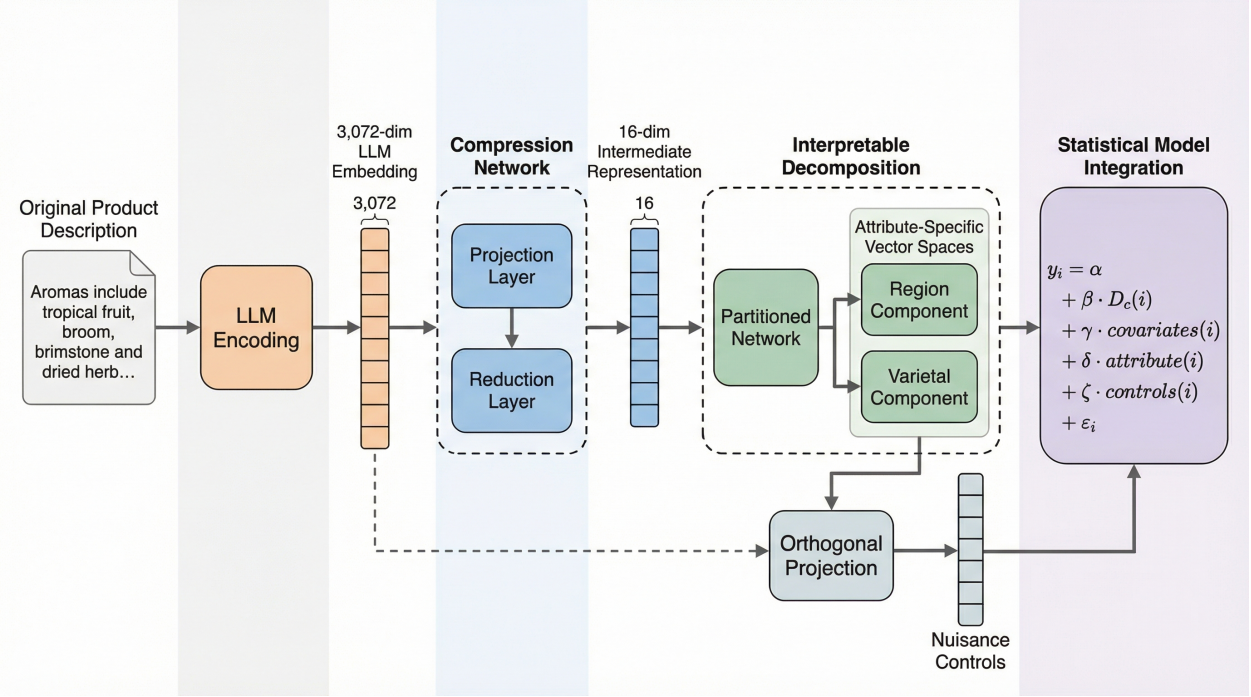


Figure 1: Schematic of BRIDGE Architecture

Note: An LLM maps an unstructured product description to a 3,072-dimensional embedding. A compression network—consisting of a projection layer and a reduction layer—compresses this embedding to produce a 16-dimensional intermediate representation. A partitioned network maps the intermediate representation to attribute-specific vector spaces (e.g., Region and Varietal components). The original LLM embedding is projected onto the orthogonal complement of the attribute-specific subspace to derive nuisance controls. Both the attribute representations and nuisance controls feed into the statistical model.

To ensure interpretability in the attribute-specific components, BRIDGE is trained using a multi-term loss function:

$$\mathcal{L} = \underbrace{\sum_{k=1}^{K} \mathcal{L}_{\text{CE}}^{(k)}}_{\text{classification}} + \lambda \underbrace{\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \log\left(1 + \sum_{j \in \mathcal{M}_i \setminus \{i\}} \exp\left(\frac{\cos(\mathbf{e}_{k,i},\, \mathbf{e}_{k,j})}{\tau}\right)\right)}_{\text{contrastive}} \tag{3}$$

where K is the number of focal attributes; $\mathcal{L}_{\text{CE}}^{(k)}$ is the cross-entropy classification loss for attribute k; $\mathbf{e}_{k,i}$ is the ℓ_2-normalized representation for attribute k of stimulus i; $\mathcal{M}_i$ denotes the mini-batch containing stimulus i; τ is a temperature parameter; and λ is the contrastive weight (the `bridge`

package supports automatic tuning of these values via Optuna; see Web Appendix §F.2).

The classification term focuses on accurately predicting the attributes (e.g., region, varietal) associated with each output node. The contrastive term serves two complementary purposes: it encourages the attribute-specific vector spaces to be distinct from one another (between-attribute distinctiveness), and encourages the representations of distinct levels within each attribute to be well-separated (within-attribute expansiveness). The loss function balances interpretability and performance, as the classification term benefits from the interpretable layer being only sufficiently large to express the key attribute-specific information in the product descriptions—a larger layer can lead to less precise training and lower performance on validation loss—whereas the contrastive term benefits from larger dimensional spaces, as these facilitate orthogonality in the attribute- and attribute-level-specific representations.

Statistical controls for the nuisance variables are developed by projecting the original LLM representations onto the orthogonal complement of the subspace spanned by the attribute-specific representations. The orthogonal complement of any subspace W is the subspace of vectors orthogonal to every vector in W. This ensures that the nuisance controls are, by construction, orthogonal to the attribute representations.

Step-by-Step Procedure for Implementing the Research Design

Conducting a study using BRIDGE involves five key steps. As depicted in Figure 2 (for a detailed discussion on each step, see Web Appendix §B), *Step 1* is collecting product descriptions. These can be sourced from existing datasets, collected from e-commerce platforms, or generated using AI models. *Step 2* involves participant data collection. BRIDGE supports within-subjects, between-subjects, and mixed experimental designs. It (a) converts the unstructured product descriptions to high-dimensional numerical representations ("embeddings") in *Step 3* and (b) refines these representations into a partitioned neural network, which generates structured, interpretable, and low-dimensional representations in *Step 4*. BRIDGE maps each attribute of interest to a

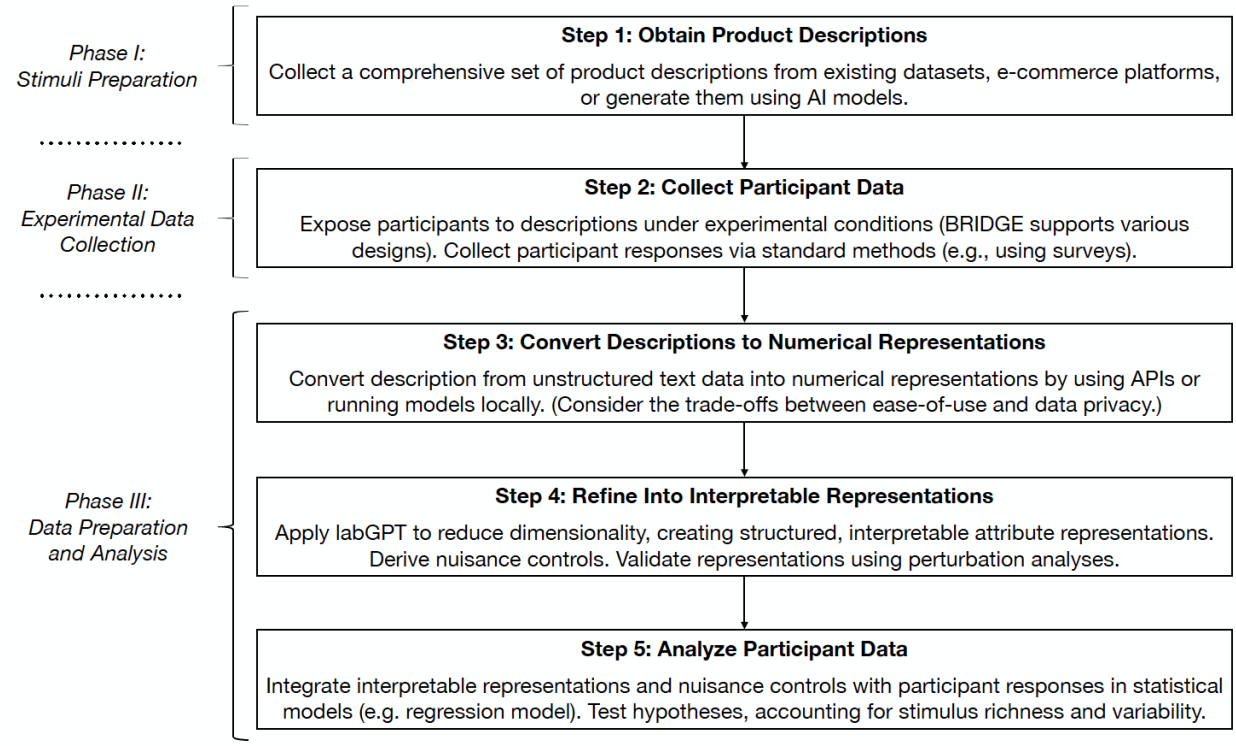


Figure 2: Procedure of a Study Using BRIDGE

distinct numerical component, facilitating clear comparisons across diverse stimuli.[5] BRIDGE also derives statistical controls for nuisance variables by orthogonalizing intermediate representations, ensuring extraneous textual variations are controlled. BRIDGE thus produces two outputs: (i) low-dimensional, attribute-specific representations that capture how focal constructs are conveyed in text, and (ii) orthogonal nuisance controls that absorb residual, non-focal textual variation. These outputs serve different inferential roles downstream—nuisance controls adjust for unobserved textual confounding, while the attribute representations capture within-attribute variation in how an attribute is expressed. Before proceeding to analysis, researchers should validate the trained model by reporting classification accuracy and weighted F1 for each focal attribute on a held-out test set, confirming that the learned representations preserve the intended attribute distinctions. In *Step 5*, researchers leverage the refined attribute representations to analyze participant responses.

[5]Representations derived from BRIDGE (e.g., for region and varietal in our wine study) can be viewed as pre-trained embeddings. Researchers working with the same stimuli could potentially leverage these existing representations directly, bypassing the need for retraining.

Validation: Perturbation Analyses

To evaluate whether the extracted representations in *Step 4*—and the downstream statistical conclusions in *Step 5*—are robust to nuisance variation, BRIDGE introduces a validation technique termed *perturbation analyses*. The objective is twofold: to ensure that the model's focal attribute representations remain stable with the inclusion of confounding variation, and to verify that its nuisance controls effectively absorb any introduced noise.

The procedure involves introducing controlled modifications (perturbations) to the stimuli that alter non-focal elements without changing focal attributes. For example, this might involve appending marketing phrases (e.g., "New Design!", "Limited Stock!") to an existing product description to deliberately introduce confounding nuisance variation. These perturbed descriptions are then processed by the trained BRIDGE model and used in the downstream econometric model.

Evaluation proceeds in two steps. First, representational stability is assessed by calculating similarity metrics (e.g., the RV coefficient) between the original and perturbed outputs. Robustness is indicated by high similarity across the attribute representations, coupled with low similarity across the nuisance controls (confirming the controls captured the injected noise). Second, the final statistical analysis is replicated using the perturbed representations. If the resulting coefficient estimates, significance levels, and model fit align with the original findings, it confirms that the observed effects stem from genuine theoretical constructs, as the deliberately introduced confounders, and the alternative explanations that they present, are effectively absorbed by the model. Full implementation details and code are provided in Web Appendix §C.

Enhancing Research Transparency, Objectivity, and Reproducibility

BRIDGE enhances research transparency, objectivity, and reproducibility by mapping the research process into a structured, *end-to-end algorithmic pipeline* that can be precisely pre-specified and pre-registered before data collection and analysis. Traditional experimental designs often require numerous subjective decisions about stimulus selection, simplification, and data analysis—decisions that can introduce researcher degrees of freedom, unintentionally compromising reproducibility

and credibility (the "garden of forking paths" problem, Gelman and Loken 2013). BRIDGE allows researchers to commit *a priori* to: (1) a *real-world dataset* and a *stimulus sampling strategy* (e.g., random sampling) to reduce stimulus-selection bias, (2) an *embedding method* (e.g., an open-source model such as NV-Embed-v2) to encode the stimuli, (3) a *BRIDGE architecture* or a *systematic procedure* for determining it (e.g., using automated tools like Optuna within defined parameters, as shown in Web Appendix §F.2), and (4) a complete *analytical plan*, including independent and dependent variables, derived representations, nuisance controls, and models for hypothesis testing. Leveraging a complete real-world dataset and executing a pre-defined, pre-registered framework reduces subjectivity and minimizes researcher degrees of freedom related to stimulus selection, data processing, and analysis. Our detailed researcher's guide and accompanying code (Web Appendix §B) are designed to facilitate such transparent and reproducible workflows. A fully reproducible example based on our Coffee Certification Experiments—including pipeline code, experiment data, and pre-computed outputs—is provided alongside the `bridge` package.

MONTE CARLO STUDIES EVALUATING BRIDGE'S PERFORMANCE

We first apply BRIDGE to synthetic data under controlled conditions, which provide a clear ground truth of treatment effect and confound structure; this study allows us to validate its inferential performance. The complete study design and Python code are presented in Web Appendix §D; below, we provide an overview and key results.

Confounds can threaten inference in two ways. First, when the confound co-varies with the treatment and directly shifts preferences (Kirk 2013), it creates an omitted variable bias on the main effect of the treatment. We term this channel, *main effect confounding*. Second, when it changes the *effective strength* of a focal attribute, it acts as an unobserved moderator (Blackwell and Olson 2022). In this case, it creates an omitted variable bias on the interaction of the treatment with the attribute. We term this channel, *interaction effect confounding*.

The implications for theory development are complex as these channels can elevate both Type I and Type II errors. If there is no treatment effect, a bias can lead to a false positive (a Type I

error). If the bias attenuates the estimate, it can lead to a false negative (a Type II error).

Simulation Study Design

Consider an experiment investigating whether exposure to one environmental concern (factory-farmed meat) shifts consumer preferences towards another environmental feature (organic farming). Participants are randomly assigned to a treatment condition (information about the detrimental effects of factory farming) or a control condition (no such information). They then evaluate 24 packaged salads. Each salad is described by a product description that conveys its attributes—organic status, size, type, and weight—in natural language. The researcher's goal is to estimate the causal effect of the information about factory farming on preferences for a meat-free meal option, and whether it spills over to the organic premium.

Crucially, the study aims to use real-world descriptions to address the stimulus sampling problem and boost generalizability. Furthermore, the real-world product descriptions obtained by the researcher vary in elaboration: some are terse and factual, others are vivid and creative.

This variation can create a confound. For instance, if the treatment-condition participants were to encounter more elaborately written descriptions, a naive analysis would not be able to discriminate between the treatment effect and the utility/disutility from elaboration. In addition, elaboration may also amplify how much consumers value the organic attribute: a vividly described organic salad may command a larger premium than a tersely described one. Consequently, not only may the analysis yield a positive bias on the treatment effect (and therefore a false positive, a Type I error) but also a positive bias on organic interacted with the treatment (also a false positive, another Type I error).

Therefore, the objective is to recover the true treatment effect and the true interaction effect without observing the confound (elaboration), using only the stimuli themselves. The five estimators we evaluate represent a hierarchy of approaches to this challenge, from ignoring the confound entirely (No Controls) to exploiting privileged knowledge of the style assignments (Oracle). The question is whether BRIDGE, which learns nuisance controls and attribute representations from

the descriptions, can match the infeasible Oracle.

Data-Generating Process (DGP)

We simulate data for $N = 1{,}000$ participants, randomly assigned to a treatment or control condition, each completing 24 evaluation tasks on a 13-point Likert scale. To introduce a confound, we treat product descriptions as differing in elaboration across three styles—Factual (least elaborate), Engaging (moderately elaborate), and Creative (most elaborate). We present participants in the treatment condition with a greater proportion of Engaging and Creative descriptions (8 Engaging, 12 Creative, and 4 Factual per participant), and the control group with more Factual descriptions (12 Factual, 8 Engaging, and 4 Creative per participant); these differences in style allocation are unobserved.

The Oracle estimator is efficient but uses knowledge of the unobserved confound (description style). It is therefore infeasible. Feasible estimators use only the observed information (product descriptions, product attributes). The objective of the simulation is to establish the efficacy of BRIDGE as a feasible estimator, to match both the known ground truth in the estimation and the estimates from Oracle estimator.

Formally, product j is characterized by the tuple $\Big(\text{Organic}_j, \text{Size}_j, \text{Type}_j, \text{Weight}_j, \text{Treatment}_j, \text{Style}_j\Big)$, where Organic_j and Treatment_j are binary indicators, Size_j and Type_j are categorical, Weight_j is continuous, and $\text{Style}_j \in \{\text{Factual}, \text{Engaging}, \text{Creative}\}$. $\text{Intensity}_j \in \{0, 1, 2\}$, maps the three styles to increasing levels of descriptive elaboration. Participant i's latent utility for product j is:

$$\begin{aligned}\text{preference}_{ij} = \alpha + \beta_1 \text{Treatment}_j + \beta_2\Big(1 + \lambda \cdot \text{Intensity}_j\Big) \cdot \text{Organic}_j + \beta_3(\text{Treatment}_j \times \text{Organic}_j) \\ + \beta_4 \text{Size}_j + \beta_5 \text{Type}_j + \beta_6 \text{Weight}_j + \theta \cdot (\text{Intensity}_j - 1) + \varepsilon_{ij},\end{aligned}$$

where $\beta_1 = 0.25$, $\beta_2 = 0.50$, and $\beta_3 = 0.50$ capture the treatment effect, organic premium, and their interaction (spillover), respectively; β_4, β_5, and β_6 represent size, type, and weight effects (with $\beta_6 = 0.10$); and $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$.

The parameter θ controls the *additive* effect of style on preferences (main effect confounding),

shifting utility by $-\theta$, 0, and $+\theta$ for Factual, Engaging, and Creative descriptions, respectively. The parameter λ controls the *moderating effect* of description intensity on the organic premium (interaction effect confounding). When $\lambda > 0$, the organic effect becomes style-specific: β_2 for Factual, $\beta_2(1 + \lambda)$ for Engaging, and $\beta_2(1 + 2\lambda)$ for Creative.

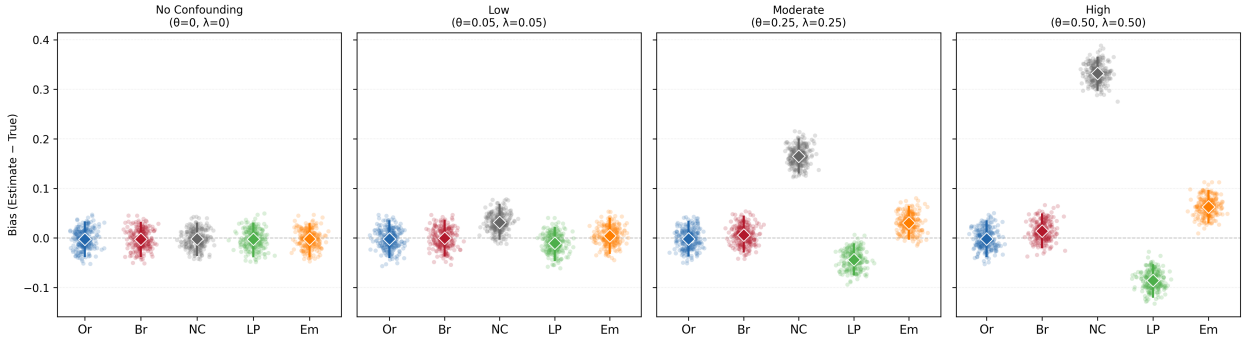
Analytical Approaches

We evaluate five estimators:

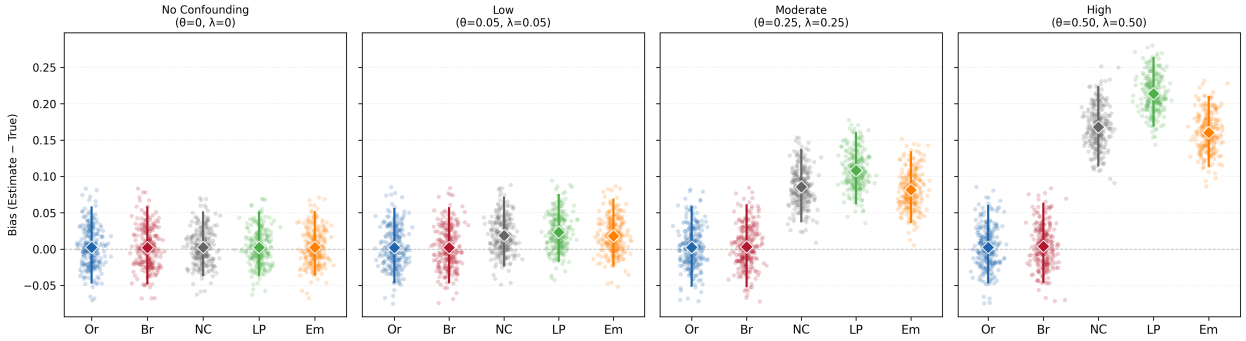
1. **Oracle (including Style Dummies):** An infeasible estimator that includes dummies describing the unobserved confound structure alongside style $\times$ organic interaction terms.
2. **BRIDGE:** Interacts BRIDGE's orthogonal nuisance controls and attribute representations with the organic indicator to capture both main effect and interaction effect confounding, without observing style.
3. **No Controls:** Omits text controls entirely; assumes there is no confound.
4. **Linear Projection:** A feasible estimator that includes nuisance controls derived by using Ridge regression on the attribute labels and then applying singular value decomposition (SVD) to the residuals—a linear decomposition of the original LLM embedding.
5. **Empath Controls:** A feasible estimator that includes control variables derived from the Empath psycholinguistic library (Fast, Chen, and Bernstein 2016), which uses deep learning (neural embeddings trained on 1.8 billion words) to generate pre-validated measures.

Results

We examine parameter vectors corresponding to four scenarios (no confounding: $\theta = 0$, $\lambda = 0$; low confounding: $\theta = 0.05$, $\lambda = 0.05$; moderate confounding: $\theta = 0.25$, $\lambda = 0.25$; high confounding: $\theta = 0.5$, $\lambda = 0.5$). Figure 3 describes the estimated treatment effect and interaction effect (treatment $\times$ organic) for 250 different simulations (where we vary the seed). In each subplot, a dot represents an estimate for one random seed, the diamond indicates the mean bias, and vertical bars span the 2.5th–97.5th percentile interval. The dashed line at zero indicates no bias.



(a) Treatment effect ($\beta_1 = 0.25$)



(b) Treatment × Organic interaction ($\beta_3 = 0.50$)

Figure 3: Estimation bias (estimate − true value) across four confounding levels: no confounding ($\theta = 0$, $\lambda = 0$), low ($\theta = 0.05$, $\lambda = 0.05$), moderate ($\theta = 0.25$, $\lambda = 0.25$), and high ($\theta = 0.50$, $\lambda = 0.50$). Each dot represents one of 250 random seeds; diamonds indicate the mean bias; vertical bars span the 2.5th–97.5th percentile interval. The dashed line at zero indicates no bias. Or = Oracle (infeasible), Br = BRIDGE, NC = No Controls, LP = Linear Projection, Em = Empath Controls.

BRIDGE closely approximates the Oracle—the efficient but infeasible estimator—across all four scenarios (Figure 3). For both the treatment effect and the treatment × organic interaction, the BRIDGE and Oracle bias distributions are centered on zero and are virtually indistinguishable, regardless of confounding intensity. The same holds for the three style-specific organic effects, where BRIDGE and Oracle are again virtually identical across all confounding levels (see Web Appendix §D for the corresponding figures). In contrast, the three alternatives align with the true values only in the absence of confounding ($\theta = 0$, $\lambda = 0$). Even at low confounding ($\theta = 0.05$, $\lambda = 0.05$), the No Controls estimator shows substantial treatment bias. At moderate confounding ($\theta = 0.25$, $\lambda = 0.25$), all three alternatives are severely biased on the interaction effect and, to

varying degrees, on the treatment effect. These biases are reflected in coverage rates: an accurate estimator's 95% confidence interval should contain the true value in 95% of the simulations (237.5 of 250). With moderate confounding, No Controls, Linear Projection, and Empath's coverage of the true treatment effect drops to 0, 99, and 179, respectively, while coverage of the true interaction effect to just 26, 4, and 30. In contrast, BRIDGE and Oracle closely match the expected rate on both the treatment effect (240 and 244) and the interaction effect (236 and 237).

Table 1: Monte Carlo Simulation Results ($\theta = 0.25$, $\lambda = 0.25$)

	Oracle	**BRIDGE**	**No Controls**	**Linear Projection**	**Empath Controls**
Treatment ($\beta_1 = 0.25$)	0.240 (0.021) [0.199, 0.280] ✓	0.248 (0.021) [0.207, 0.288] ✓	0.428 (0.019) [0.390, 0.466] ✗	0.208 (0.020) [0.169, 0.248] ✗	0.292 (0.020) [0.253, 0.330] ✗
Treatment × Organic ($\beta_3 = 0.50$)	0.512 (0.029) [0.454, 0.569] ✓	0.512 (0.029) [0.455, 0.569] ✓	0.567 (0.027) [0.513, 0.621] ✗	0.593 (0.027) [0.540, 0.645] ✗	0.563 (0.027) [0.510, 0.616] ✗
Organic, Factual ($\beta_2 = 0.50$)	0.518 (0.024) [0.470, 0.566] ✓	0.518 (0.024) [0.470, 0.565] ✓	0.577 (0.019)† [0.539, 0.615] ✗†	0.565 (0.019)† [0.528, 0.602] ✗†	0.588 (0.019)† [0.550, 0.626] ✗†
Organic, Engaging (0.625)	0.611 (0.027) [0.557, 0.664] ✓	0.608 (0.027) [0.556, 0.661] ✓	—	—	—
Organic, Creative (0.75)	0.688 (0.032) [0.626, 0.751] ✓	0.689 (0.031) [0.628, 0.750] ✓	—	—	—
R-squared	0.481	0.481	0.452	0.479	0.470

Seed 118. Standard errors in parentheses, 95% confidence intervals in brackets.
✓ indicates ground truth falls within the 95% CI, ✗ indicates it does not.
†Pooled organic coefficient; these estimators cannot decompose by style.
Style-specific organic effects for BRIDGE and Oracle computed via the delta method.
All models include size and type dummies.

Table 1 presents coefficient estimates and 95% confidence intervals for all estimators in the moderate confounding condition ($\theta = 0.25$, $\lambda = 0.25$). The Oracle, which observes the unobserved style assignments, recovers all five focal coefficients: treatment ($\hat{\beta}_1 = 0.240$, CI $[0.199, 0.280]$), the treatment × organic interaction ($\hat{\beta}_3 = 0.512$, CI $[0.454, 0.569]$), and style-specific organic effects of 0.518, 0.611, and 0.688 for Factual, Engaging, and Creative (true values: 0.50, 0.625, 0.75). This privileged information eliminates confound-related uncertainty, but it is unavailable to researchers in practice: style assignments are unobserved. The Oracle thus serves as a benchmark for the best achievable inference.

BRIDGE, without observing style, approaches the Oracle. The treatment effect ($\hat{\beta}_1 = 0.248$, CI

$[0.207, 0.288]$) and the treatment $\times$ organic interaction ($\hat{\beta}_3 = 0.512$, CI $[0.455, 0.569]$) are both consistent with the ground truth. The style-specific organic effects—0.518, 0.608, and 0.689—are virtually indistinguishable from the Oracle. The mechanism is the interaction of BRIDGE's learned representations with the organic indicator: both the nuisance controls and the attribute representations carry style-related information that differs systematically between organic and conventional descriptions, serving as continuous proxies for the unobserved style $\times$ attribute interaction. Web Appendix Table~WA1 reports the full coefficient vector. To recover style-specific organic effects, we use the delta method: for each style s, we project $\hat{\beta}_2$ to the conditional mean nuisance and attribute representation profile for style s, with standard errors computed from the gradient and the relevant covariance submatrix.

The absence of any text controls leads to severe bias. The No Controls estimator overestimates the treatment effect by 71% ($\hat{\beta}_1 = 0.428$ versus the true 0.25) because the main effect confound is erroneously attributed to treatment. The treatment $\times$ organic interaction is similarly inflated ($\hat{\beta}_3 = 0.567$ versus the true 0.50), reflecting interaction effect confounding.

Linear Projection underestimates the treatment effect ($\hat{\beta}_1 = 0.208$, CI $[0.169, 0.248]$, excluding the true 0.25) and fails on the interaction ($\hat{\beta}_3 = 0.593$, CI $[0.540, 0.645]$, excluding the true 0.50). The linear decomposition absorbs some style variance but cannot fully capture the relationship between style and attribute effects.

Empath lexical features overestimate the treatment effect ($\hat{\beta}_1 = 0.292$, CI $[0.253, 0.330]$, excluding the true 0.25) and fail on the interaction ($\hat{\beta}_3 = 0.563$, CI $[0.510, 0.616]$, excluding the true 0.50). More broadly, all three feasible alternatives yield a single pooled organic coefficient (~ 0.57) that collapses the style-varying effects; none can distinguish the Factual-specific effect (0.50) from the Creative-specific effect (0.75). Only BRIDGE's interactive specification—where nuisance controls and attribute representations are interacted with the organic indicator—captures both confound channels, recovering the three latent effects that the Oracle specifies with discrete dummies.

Discussion

The results demonstrate that BRIDGE recovers the full set of structural parameters—treatment, organic × treatment interaction, and style-specific organic effects—matching the infeasible Oracle across all five focal coefficients.

Each alternative estimator fails differently, illustrating the distinct confound channels at work. Under main effect confounding alone, lexical controls such as Empath may plausibly absorb the dominant style axis and recover the treatment effect. But when textual features (such as extent of elaboration or elaboration style) also moderate attribute strength (the λ channel), these methods collapse the attribute effect to a single average, yielding biased estimates of the interaction effect. In these scenarios, only our BRIDGE specification is both feasible and accurate.

These limitations of alternative approaches carry practical implication. In real-world applications, the attribute space is typically large and semantically structured—our wine study, for example, involves 426 regions and 708 varietals. More fundamentally, consumer decisions likely exhibit interaction effect confounding: a terse mention of "Burgundy" and an evocative paragraph about Burgundian terroir both indicate the same region, but likely differentially activate consumer associations and preferences.

These considerations delineate a practical complexity boundary. When product descriptions are simple and few, conventional designs with carefully controlled stimuli likely suffice and algorithmic nuisance controls add little incremental value. BRIDGE becomes necessary in product contexts where many varied descriptions are the norm in the real-world marketplace: in such information-rich environments, it is only a joint specification of learned attribute representations and orthogonal nuisance controls that captures both confound channels successfully, providing a feasible approach to unbiased inference.

Because the confound structure is unknown ex ante and real-world descriptions plausibly induce confounding, BRIDGE remains the robust default. It matches simpler methods when they suffice (as the statistical controls are by design orthogonal to the product's attributes, including these controls even when the nuisance variables play a minimal role does not affect consistency,

as illustrated in Figure 3). And it is the only feasible and accurate approach that can handle both channels of confounding. The empirical studies that follow further validate BRIDGE: controlled coffee experiments establish the efficacy of BRIDGE's nuisance controls, while a large-scale wine study applies BRIDGE in a high-dimensional attribute space.

COFFEE CERTIFICATION EXPERIMENTS

Having established BRIDGE's performance in controlled Monte Carlo simulations with specified confound structure, we test BRIDGE in experiments with human participants. Researchers often do not observe whether (and which) non-focal features of the stimuli co-vary with the focal treatment–the confound structure may not be known in advance. Even when a specific confound is suspected, accounting for it typically requires creating additional stimulus versions, running follow-up studies, or collecting more measures (for covariate adjustments). BRIDGE addresses potential confounds by extracting nuisance controls from the stimuli without explicit confound specification.

To examine the efficacy of BRIDGE, we situate our experiments in coffee certification framing: fair trade in Experiment 1 and organic in Experiment 2. Product texts that include certification information can vary in length, sometimes in systematic ways. For example, a fair trade or organic disclosure may be conveyed succinctly (e.g., a certification label only) or elaborately (e.g., sourcing and standards information) compared to products without the certification disclosure. Such variation can introduce a text-length confound that obscures the true treatment effect.

In the experiments, we deliberately introduce a text-length confound to assess BRIDGE's performance. Participants evaluate a pair of coffee descriptions and are randomly assigned to conditions in which the treatment description is shorter than, matched to, or longer than the control description. The matched condition is the standard stimulus-control design in which the stimuli differ only on the focal dimension, approximating an "unconfounded" treatment effect; in the remaining conditions, the treatment effect is confounded by description length. We then examine whether BRIDGE's nuisance controls (which are constructed without explicitly specifying

length as a confound) recover treatment-effect estimates comparable to the matched benchmark and how it compares with other conventional estimators.

Method

Participants and design.

Participants were recruited from Prolific U.S. panel, who completed these mutually exclusive studies for a small monetary compensation. In each experiment, they were randomly assigned to one of the three between-subjects conditions (treatment framing length: shorter-than-control, length-matched, longer-than-control). Experiment 1 recruited 353 participants (43.9% women; 55.5% men; 0.6% non-binary; M_{age} = 44.9; https://aspredicted.org/7dt3ev.pdf). Experiment 2 recruited 352 participants (49.7% women; 48.9% men; 1.1% non-binary; 0.3% prefer not to say; M_{age} = 44.7; https://aspredicted.org/fz6y8h.pdf).

Procedure and measures.

Participants evaluated a pair of coffee descriptions—one with certified framing (treatment; Experiment 1: fair trade, Experiment 2: organic) and one without (control)—and indicated their relative preference on a 7-point scale (1 = strongly prefer Option A, 7 = strongly prefer Option B). All participants were presented with the same two underlying coffee profiles (nutty/balanced; chocolatey/full-bodied). For each participant, certified framing was randomly assigned to one profile, and the left/right placement of the options was randomized. We varied the treatment description length (short, medium, or long); the control description was always medium length. This design yields a length-matched benchmark condition (medium treatment, medium control), and provides a controlled setting to assess how experimentally induced "nuisance" variation in text length (shorter-than-control, length-matched, longer-than-control) can affect inference about the focal treatment effect. Finally, participants provided basic background information (gender, age). Detailed stimulus materials and methods are available in Web Appendix §E.

Model Specifications.

The key dependent variable is participants' 7-point recoded preference, ranging from -3 (strongly prefer the Base Option [control option]) to 3 (strongly prefer the Comparison Option [certification-framed option]). We fit two Bayesian linear regressions: the Naïve and Proposed models. The models are designed to test, using different approaches, the extent to which participants' preferences are influenced by the certified frame (treatment) when its length is also varied (confound).

Specification 1: Naïve Model As in our preregistrations, the first model estimates the effect of treatment frame on recoded preference using a Bayesian linear regression with condition indicators. The treatment/medium-length condition served as the reference category.

In Experiment 1, the utility that participant i assigned to the Comparison Option was represented as:

$$u_i = \beta_0 + \beta_1 \mathbb{I}(\text{FairTrade+Short}_i) + \beta_2 \mathbb{I}(\text{ FairTrade+Long }_i) + \varepsilon_i. \tag{4}$$

where u_i represents the utility of participant i for the Comparison Option; β_0 is the intercept, representing the effect of fair-trade/medium-length option; FairTrade+Short_i and FairTrade+Long_i are the dummy variables for the corresponding conditions; β_1 and β_2 are the coefficients capturing the influence of shorter or longer description length on participants' preference, and; ε_i is an i.i.d. Gaussian error term.

A significant intercept β_0 indicates that the fair-trade/medium frame differs significantly from the non-fair trade/medium control. Significant β_1 or β_2 coefficients indicate additional effects of the shorter or longer fair-trade frame relative to the fair-trade/medium frame.

In Experiment 2, a parallel utility function was specified for certified organic framing. We estimated each experiment separately; results are reported in Web Appendix §E.

Specification 2: BRIDGE This model adds a vector of nuisance controls generated by our BRIDGE method to estimate the utility that participant i assigned to the Comparison Option. These controls are extracted from the orthogonalized residuals of the LLM embeddings relative

to the learned focal attribute representations. BRIDGE does not require (1) knowledge of the experimental conditions nor (2) specification of what the confound is; it learns the nuisance structure directly from the text. Thus, the estimator is both feasible and flexible.

Specifically, we represent utility as:

$$u_i = \beta_0 + \mathbf{z}_i^\top \boldsymbol{\gamma} + \varepsilon_i. \tag{5}$$

where u_i represents the utility of participant i for the Comparison Option; β_0 is the intercept, representing the effect of certified/medium-length option; $\mathbf{z}_i$ is a $K \times 1$ vector of nuisance controls for participant i, constructed using the BRIDGE method; $\boldsymbol{\gamma}$ is a $K \times 1$ vector of coefficients corresponding to the nuisance controls, and; ε_i is an i.i.d. Gaussian error term.

Additional Specification: Word-Count Model A model that tests the extent to which adding word count as a variable controls for the specific confound (length) in measuring β_0 (the effect of treatment frame).

Results: Relative preference for certified coffee

All completed observations were included in our analyses. Experiments 1 and 2's key findings from the Naïve and the Proposed models are described in Table 2. Detailed methods, stimuli, and model specifications are provided in Web Appendix §E.

Naïve Model

Given that certification can be conveyed with varying elaboration, we first report the naïve treatment effect implied by each implementation (shorter, length-matched, and longer). These correspond to the estimates a conventional study would obtain if it used only that version of the treatment text. As shown in Table 2, when description length is matched (both medium-length), in both experiments we observe that treatment framing increases the 7-point recoded preference, by +1.32 in Experiment 1 (fair trade) and +.97 in Experiment 2 (organic). This estimate serves

Table 2: Coffee Certification Experiments

	Experiment 1: Fair Trade		Experiment 2: Organic	
Estimator	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Naive (matched)	1.32	[0.96, 1.69]	0.97	[0.56, 1.38]
Naive (shorter)	0.16	[−0.22, 0.52]	0.38	[−0.02, 0.78]
Naive (longer)	1.29	[0.88, 1.69]	1.26	[0.89, 1.61]
Naive (pooled)	0.92	[0.70, 1.14]	0.88	[0.66, 1.11]
Word Count	0.80	[0.57, 1.04]	0.76	[0.51, 0.99]
BRIDGE	1.24	[1.00, 1.49]	1.02	[0.51, 1.54]

Note. Naive (matched/shorter/longer) estimates are derived from a single Bayesian linear regression with condition indicators (medium-length as reference): matched $= \hat{\beta}_0$, shorter $= \hat{\beta}_0 + \hat{\beta}_1$, longer $= \hat{\beta}_0 + \hat{\beta}_2$. Naive (pooled) is the weighted average of the condition-specific effects, with weights proportional to sample sizes. Word Count controls for the word-count difference Δ_{wc} between treatment and control descriptions. BRIDGE uses one nuisance control extracted from AI text embeddings; results with two nuisance controls are similar. All models use Gaussian errors and are estimated separately by experiment. All 95% CIs are posterior credible intervals.

as our key benchmark: it holds constant the confound (description length) while keeping the descriptions comparable (with residuals randomized across participants), corresponding to the standard stimulus-control approach. However, designing perfectly controlled stimuli may be difficult or not always possible. When the treatment text is shorter than the control text, the estimated effect of treatment framing reduces to +.16 in Experiment 1 and +.38 in Experiment 2; in each respective experiment, the shorter-than-control estimate is substantially smaller than the length-matched estimate and not within its 95% CI. When it is longer than the control text, the estimated effect is +1.29 in Experiment 1 and +1.26 in Experiment 2, both comparable to the respective length-matched benchmark (within its 95% CI).

We lay out these three scenarios to illustrate how the presence of a confound can affect inference using conventional approaches. These estimates reflect situations in which the confound (description length) is observed. When the confound is mixed and unobserved (e.g., when a researcher is unaware or ignores length differences by pooling across scenarios), the naïve pooled estimate is +0.92 and +0.88 in Experiments 1 and 2, respectively. In Experiment 1, uncontrolled variation in description length attenuates the effect of fair-trade framing (about 30% smaller)

relative to the length-matched benchmark; in Experiment 2, the pooled estimate is comparable to the organic benchmark. Uncontrolled nuisance variations in experimental texts can lead to materially different conclusions (e.g., underestimating the treatment effect, increasing the risk of a false negative). Thus, even in a randomized experiment, differences in treatment versus control text length can skew inferences about a focal treatment effect embedded in the stimulus text.

BRIDGE

In Experiment 1, BRIDGE ($\hat{\beta}_0 = 1.24$, 95% CI [1.00, 1.49]) recovers the treatment effect comparable to the length-matched benchmark ($\hat{\beta}_0 = 1.32$; posterior difference $\delta = -0.081$, 95% CI $[-0.528, 0.361]$). In Experiment 2, BRIDGE ($\hat{\beta}_0 = 1.02$, 95% CI [0.51, 1.54]) similarly recovers the benchmark ($\hat{\beta}_0 = 0.97$; $\delta = .051$, 95% CI $[-0.626, 0.707]$). In both experiments, BRIDGE effectively recovers the unconfounded treatment effect without knowledge of the experimental conditions. The results also suggest a stronger certification effect for fair-trade framing relative to organic.

Word-Count Model

If the researcher is aware of the text-length confound, one possibility is to treat it as a covariate and statistically control for it. We fit a linear model controlling for the word count difference between the treatment and control descriptions. In Experiment 1, this yields $\hat{\beta}_0 = 0.796$ with a slope of .021 per word of difference, producing condition-specific estimates of 0.564 for the shorter condition (10 words), 0.796 for the matched condition (21 words), and 1.450 for the longer condition (52 words). In Experiment 2, the corresponding model yields $\hat{\beta}_0 = .756$ with a slope of 0.018, producing estimates of 0.508 (shorter, 10 words), 0.756 (matched, 24 words), and 1.306 (longer, 55 words). Across both experiments, the word-count model overcorrects at the matched condition; in Experiment 1, this difference (overcorrection by 0.528) is significant (95% CI $[-0.971, -0.096]$ does not contain 0). A standard word-count covariate does not adequately capture the relationship between description length and preference, mischaracterizing the treatment effect across conditions.

Coffee Certification Experiments Discussion

Experiments 1 and 2 show that confounding variation in text length can distort the estimated effect of certified framing under conventional analyses. Conceptually, the design aligns with the Monte Carlo study but with real participants. The pooled naïve estimates are consistent with main effect confounding: uncontrolled length variation attenuates the treatment effect relative to the length-matched benchmark. The inclusion of word-count does not reliably recover the benchmark, suggesting standard covariate adjustments might not be adequate for this confounding. In contrast, even when the confound was not explicitly specified, BRIDGE recovers treatment effects comparable to the length-matched benchmark in the respective experiments. This motivates the use of BRIDGE's nuisance control, as it helps isolate the treatment effect while flexibly accounting for nuisance variation in text. Broadly, the experiments show that BRIDGE can complement careful stimulus design by providing a safeguard when residual or unanticipated confounding variation remains.

PREFERENCE DYNAMICS IN UNSTRUCTURED REAL-WORLD PRODUCT DESCRIPTIONS

Next, we demonstrate how BRIDGE can (1) enable the sampling of myriad unaltered real-world product texts as study stimuli, addressing the stimulus-sampling problem, and (2) apply to hypothesis-testing in information-rich environments.

Normative theories view consumer preferences as stable and merely revealed during decision making (Rabin 1998). Mounting evidence, however, shows that preferences are often constructed based on how options are presented (Payne et al. 2000; Slovic 1995). Making repeated choices can stabilize preference as consumers learn their preferences (Amir and Levav 2008; Hoeffler and Ariely 1999). The findings align with anchoring—individuals initially anchor on accessible numerical information as a reference point and, with cognitive effort, adjust closer to or farther from that anchor (Ariely, Loewenstein, and Prelec 2003; Spicer et al. 2022). Much anchoring evidence relies on numbers, with some evidence extending to semantic words (Chernev 2011). Previous studies on preference dynamics employ structured choice contexts in which options are

defined by two simpler, aligned attributes (Amir and Levav 2008; Donkers et al. 2020).

However, consumers in real-world settings often encounter options described with unstructured text—wine tasting notes and coffee flavor narratives—rich with subjective features that may not align across options. Learning from earlier choices is more difficult in such environments because (1) the consequences of a choice can be hard to trace back to specific antecedents (Einhorn and Hogarth 1981), and (2) the concreteness of the information can affect its meaning and implications for decision making (Ebbesen and Konecni 1980; Martin, Seta, and Crelia 1990), often requiring more cognitive effort (Stone and Schkade 1991). Therefore, some scholars (Ebbesen and Konecni 1980; Einhorn and Hogarth 1981; Gigerenzer 1991) argue that judgment phenomena observed with simpler, stylized stimuli may not occur in information-rich environments; it remains unclear whether initial product descriptions affect subsequent decisions.

We examine how consumers develop preferences when making sequential choices from real-world, unstructured product texts. In two experiments, participants repeatedly viewed pairs of products described in prose. They are asked to choose their preferred option (our wine study) or indicate a relative preference (our coffee anchoring study). For each participant, each product in the sequence is drawn randomly (without replacement) from 119,955 wines or 36 coffees.

We posit that preferences for complex, verbally described products are initially constructed but become largely invariant, exhibiting consistency in subsequent choices. When consumers encounter products described in prose, they need to devote effort to assess this information. Product options encountered at the outset can activate the accessibility of select semantic knowledge about the options (Strack and Mussweiler 1997) and set standards of comparison (Mussweiler 2003). Because an anchor's influence grows when people actively read or think about the anchor (Chapman and Johnson 1999), the initial products provide a reference frame for consumers to discover their preference structure (i.e., attribute trade-off weights), thereby aligning subsequent preferences with the initial, incidental product anchors.

To test this hypothesis, we present each participant with a sequence of choice tasks involving pairs of randomly selected, real-world product descriptions in prose. We design the experiment so

that (1) the initial products (anchors) are randomly chosen and therefore incidental; (2) subsequent products are also randomly chosen, ensuring exogenous variation in the distance between these subsequent products and the anchors; and (3) the real-world product descriptions are unaltered (unabridged and non-stylized), where their characteristics might not align.

We illustrate the use of BRIDGE to analyze participants' decisions. Such an analysis would be intractable using conventional methods like ANOVA because no two participants face the exact same choice task (effectively placing each participant in a unique condition) and because of the nuisance variables inherent in real-world descriptions. To provide converging evidence, we then conduct a follow-up experiment (Web Appendix §G) using a limited number of controlled stimuli. Although this approach foregoes the benefits of broad stimulus sampling, it yields data that are amenable to traditional analysis techniques, ensuring our results reflect the phenomenon itself rather than method artifacts.

Context and Product Descriptions

Wine descriptions in the real-world marketplace are rich in sensory detail and nuance—the complexity makes wine an ideal context for examining our hypothesis of anchoring. As an initial step (Step 1 in Figure 2), we obtain a comprehensive dataset containing the verbal descriptions (tasting notes) of 119,955 wines from Wine Enthusiast, a globally recognized publication.

These tasting notes are developed through extensive blind taste tests, where professional tasters describe each wine's character and consumption experience without access to identifying labels such as the producer, name, varietal, or region. They are essential for communicating the nuanced profiles of wines to consumers and are often adapted by retailers into point-of-sale marketing materials. We incorporated this extensive set of real-world wine descriptions in our experiment, facilitated by our proposed methodology. This aligns the study materials more closely with what consumers encounter in real-world retail environments. By sampling a large, diverse corpus of real-world product descriptions, the observed findings are also more likely to generalize to other real-world wine choices.

Our conceptualization is that the activation of salient concepts in initial products can serve as a semantic reference frame (Mussweiler 2003; Strack and Mussweiler 1997). A wine's taste, cultural roots, and geographical-climatic provenance are shaped by its region of origin (terrior) and grape varietal (MacNeil 2015). Retailers often use these attributes to organize and display their products (e.g., Wine.com, a large US-based online wine shop, as well as many online or bricks-and-mortar stores). As participants are likely to draw on taste expectations associated with region and varietal when forming preferences (Latour and Latour 2010; Rocklage, Rucker, and Nordgren 2021), we focus our analysis on these attributes.

Our final dataset comprises descriptions of 119,955 wines from 426 wine-growing regions and 708 wine-grape varietals—a vast product space. Each record includes a wine's name, region, varietal, and tasting note (averaging 53 words in length; SD = 11.86 words). BRIDGE fit the data well: it achieved 98.2% classification accuracy for region (426 classes) and 98.0% for varietal (708 classes), far exceeding chance baselines of 0.23% and 0.14%, respectively. Detailed classification metrics including weighted F1 scores are reported in Web Appendix §F.2. A sensitivity analysis examining the impact of training data size on classification performance is reported in Web Appendix §F.3.

Experimental Design and Participant Data

We collected participant data in collaboration with Qualtrics, a global leader in market research (Step 2 in Figure 2). We engaged 1,000 consumers from Australia (250 participants), New Zealand (200 participants), and the United States (550 participants), ranging in age from 25 to 89, with 50.5% women. Participants met three criteria: a minimum age of 25 years (set for ethical reasons), currently employed, and reported wine consumption of at least one glass in the prior 28 days. The majority (86.5%) reported consuming at least one glass of wine per week. We instructed our service provider to ensure that the participants' demographics were representative of the wine-consuming population in their respective countries.

Each participant completed 32 sequential tasks. In each task, participants saw a pair of wines—

each selected randomly without replacement from our corpus of 119,955 real-world product descriptions—and were asked to pick their preferred option. Each wine was described by its name and tasting note. This shows BRIDGE's flexibility, as it (a) allows participants to encounter many different options and (b) presents wines in prose, resembling how consumers encounter these products in the real-world.

We opted for a relatively long sequence of 32 decision tasks. This offers a more stringent test of our hypothesized anchoring effect: Gigerenzer (1991) argued that judgment phenomena such as anchoring may dissipate when people make repeated choices over a longer sequence. The extended sequence also allows us to examine possible anchors in different parts of the sequence.

Variable Operationalization and Model

Two options were shown in each choice task. Our dependent variable was 1 if the option on the right ("right-option") was chosen and 0 if the option on the left ("left-option") was chosen. To examine anchoring, we differenced the similarity of the right-option to a *candidate anchor* and the similarity of the left-option to the same anchor. A positive (and significant) coefficent on this "right minus left" variable would indicate that participants were more likely to choose the option on the right when it was more similar to the anchor. Based on this operationalization, we construct four independent variables examining the effect of different anchor positions:

1. *Similarity to Options in the First Task (Sim^{First}):* the anchor is the pair of options in the first task; its coefficient tests the hypothesized anchoring effect.

2. *Similarity to Options in the Previous Task (Sim^{Prev}):* the anchor is the pair of options in the immediately preceding task; its coefficient tests for the recency effect.

3. *Similarity to Options in the Next Task (Sim^{Next}):* the anchor is the pair of options in the next task. As the options in the next task are unknown at the time of decision making, its coefficient serves as a placebo test. We expect a nonsignificant coefficient, ruling out alternative explanations such as pre-existing preferences (revealed preferences).

4. *Similarity to Options in the Final Task (Sim^{Final}):* the anchor is the pair of options in the final task. Like Sim$^{\text{Next}}$, its coefficient serves as a placebo test.

To rule out alternative explanations, we also calculated similarity to the second task (Sim$^{\text{Second}}$), two tasks before the current task (Sim$^{\text{Two Before}}$), two tasks after the current task (Sim$^{\text{Two After}}$), and the one-before-the-final task (Sim$^{\text{One Before Final}}$).

We fit a hierarchical Bayesian model where the utility that participant i assigns to option k in task t is:

$$\begin{aligned} u_{ikt} = \beta_{0i} &+ \beta_{1i}\text{Sim}_{ikt}^{\text{First}} + \beta_{2i}\text{Sim}_{ikt}^{\text{One Before}} + \beta_{3i}\text{Sim}_{ikt}^{\text{One After}} + \beta_{4i}\text{Sim}_{ikt}^{\text{Final}} \\ &+ \beta_{5i}\text{Sim}_{ikt}^{\text{Second}} + \beta_{6i}\text{Sim}_{ikt}^{\text{Two Before}} + \beta_{7i}\text{Sim}_{ikt}^{\text{Two After}} + \beta_{8i}\text{Sim}_{ikt}^{\text{One Before Final}} \\ &+ \sum_{q=1}^{N} \delta_{qi}\text{Nuisance}_{qikt} + \varepsilon_{ikt}. \end{aligned}$$

ε_{ikt} is i.i.d. Gumbel, yielding a logit choice model. $N = 5$, chosen based on the elbow of the scree plot of the nuisance variables.

We estimated a sequence of nested fixed-effects models. *Model 1* includes the four focal variables (Sim$^{\text{First}}$, Sim$^{\text{Prev}}$, Sim$^{\text{Next}}$, Sim$^{\text{Final}}$). *Model 2* incorporates the four additional variables (Sim$^{\text{Second}}$, Sim$^{\text{Two Before}}$, Sim$^{\text{Two After}}$, Sim$^{\text{One Before Final}}$). *Model 3* augments Model 1 by adding the BRIDGE-derived nuisance controls (Nuisance$_q$). *Model 4* includes all eight variables and the nuisance controls. In addition, to account for individual-level differences in sensitivity to these variables, we introduce participant heterogeneity through random effects. *Model 5* builds upon Model 3, allowing for heterogeneity in the intercept, the four focal similarity variables (Sim$^{\text{First}}$, Sim$^{\text{Prev}}$, Sim$^{\text{Next}}$, Sim$^{\text{Final}}$), and the nuisance controls (δ_{qi}). *Model 6* introduces participant-level heterogeneity on *all* variables—the intercept, all eight similarity measures, and the nuisance controls.

We developed measures based on the cosine similarity of the BRIDGE-derived attribute representations of the options. Specifically, we used the sum of the cosine similarity between an option's attribute representations and the attribute representations of the two options shown in

an anchor task.[6] Higher scores indicate greater similarity between a presented option and the options in the anchor task.

Each option was selected at random (without replacement from the corpus) for each participant. Participants encounter exogenous variation in both the anchors and the options presented in the tasks. Thus, the *similarity* measures vary randomly across both participants and tasks, providing exogenous variation for estimating the anchoring effect and examining alternative explanations. This stimulus-randomization procedure ensures that no single wine (or its varietal or region) could drive the focal variable, reducing potential bias from a specific stimulus.

Results

Our dataset comprises 32,000 choice tasks involving 49,548 unique wines (41.3% of available wines) from 367 wine-growing regions and 566 grape varietals. This broad sampling ensured substantial individual exposure to diversity: on average, each participant encountered 28.6 (SD = 2.99) unique wine regions and 31.6 (SD = 3.02) unique varietals across their 32 tasks. All observations were included in our analyses (i.e., no data exclusions).

Table 3 indicates participants anchored their preferences to the wines presented in the initial task, consistently favoring wines which were similar to these initial wines in subsequent choices.[7] The coefficient for "Similarity to First Task" was positive and different from zero across all seven models with this predictor, including these six models (posterior mean ranging from 0.038 to 0.042, *p*s range from 0.002 to 0.014), and those that include only single predictors (see Web Appendix §F.4.1). This finding indicates that once the product options in the initial task established a reference frame, participants' subsequent preferences remained largely invariant and consistent

[6]The BRIDGE architecture used to generate these representations was determined using hyperparameter optimization to ensure objectivity and reproducibility (see Web Appendix §F.2 for details). Post-hoc analysis confirmed a substantial degree of practical orthogonality between the resulting 8-dimensional region and varietal representations. The median pairwise cosine similarity between average attribute vectors was 0.262 (IQR: 0.079–0.510), with over half (54.5%) of pairs exhibiting similarity below 0.3 (and 28% below 0.1), indicating they capture distinct information.

[7]Full results for single-predictor models are provided in Table WA8 in Web Appendix §F.4.1, fixed-effects models (M1–M4) with nuisance control estimates in Table WA9 in Web Appendix §F.4.2, and the fully heterogeneous models (M5 and M6) including all fixed and random effects in Table WA10 in Web Appendix §F.4.3. We report exact posterior *p*-values computed from the 4,000 MCMC draws: $p_{\text{exact}} = 2 \times (1 - p_d)$, where p_d is the proportion of posterior draws on the same side as the point estimate; no distributional assumptions are required.

with that initial anchor, even though the anchor was incidentally encountered.

In contrast, the coefficients for similarity to other candidate anchors (e.g., wines shown in the previous, next, or final tasks) and the additional similarity terms were indistinguishable from zero across all models. For "Similarity to Previous Task," the 95% credible interval includes zero, which does not support recency accounts. The same holds for "Similarity to Next Task" and "Similarity to Final Task," indicating that the observed effect of the initial (incidental) anchor is not an artifact of revealed preferences.

Model fit comparisons using the Leave-One-Out Information Criterion (LOOIC) reveal that accounting for individual differences improves explanatory power. Models 5 and 6, which incorporate participant-level heterogeneity, both outperform their fixed-effects counterparts (Models 3 and 4), with Model 6 achieving the best fit overall (LOOIC = 38,293 vs. 38,320 for Model 5).

We also estimated models incorporating the position of each task in the sequence and its interaction with the similarity to the options in the first task. The interaction term's 95% credible interval included zero, suggesting that the anchoring effect does not diminish over the experiment, ruling out fatigue as an alternative explanation.[8] These results are detailed in Web Appendix §F.4.4.

To further disentangle the mechanism underlying the effect, we decomposed the similarity-to-first-task measure into two components: similarity to the wine that the participant (i) *chose* in the first task ($\text{Sim}^{\text{Chosen First}}$) and (ii) did *not* choose ($\text{Sim}^{\text{Unchosen First}}$). If the anchoring effect reflects stable preferences, participants should prefer wines resembling their initial choice and $\text{Sim}^{\text{Chosen First}}$ should drive the effect. Instead, the results reveal that only similarity to the *unchosen* wine was significant ($\hat{\beta} = 0.072$, 95% CI [0.033, 0.110], $p < 0.001$); the similarity to the chosen wine was not ($\hat{\beta} = 0.004$, 95% CI [-0.034, 0.042], $p = 0.841$). This asymmetry is consistent with the notion that the initial, arbitrary product anchor set the reference frame for comparison, driving subsequent choices. Full results of the decomposition model are reported in Web Appendix §F.4.5.

Comparing models with and without nuisance controls (e.g., Model 1 vs. Model 3) shows

[8]We thank the review team for this suggestion.

Table 3: Key Results: Wine Study

Effect	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	0.084	0.084	0.083	0.083	0.089	0.089
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.013)
Similarity to First Task	0.039	0.041	0.038	0.039	0.038	0.042
	(0.014)	(0.014)	(0.014)	(0.014)	(0.015)	(0.016)
Similarity to Previous Task	0.017	0.017	0.015	0.016	0.015	0.016
	(0.014)	(0.014)	(0.013)	(0.014)	(0.015)	(0.015)
Similarity to Next Task	-0.007	-0.007	-0.009	-0.009	-0.010	-0.009
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Similarity to Final Task	-0.003	-0.003	-0.005	-0.004	-0.004	-0.002
	(0.014)	(0.014)	(0.014)	(0.014)	(0.015)	(0.015)
Similarity to Second Task	—	0.001	—	-0.001	—	-0.003
	—	(0.014)	—	(0.014)	—	(0.016)
Similarity to Two Tasks Before	—	-0.011	—	-0.012	—	-0.011
	—	(0.014)	—	(0.014)	—	(0.015)
Similarity to Two Tasks After	—	-0.009	—	-0.009	—	-0.013
	—	(0.014)	—	(0.013)	—	(0.015)
Similarity to Second-to-Final Task	—	0.013	—	0.012	—	0.010
	—	(0.014)	—	(0.014)	—	(0.016)
LOOIC	38766	38772	38694	38700	38320	38293
Nuisance Controls Included	No	No	Yes	Yes	Yes	Yes
Full Heterogeneity Included?	No	No	No	No	Yes	Yes

Note. Fixed-effects estimates (posterior means). Standard errors reported in parentheses. Dashes (—) indicate the variable is not included. Nuisance control fixed-effect estimates and heterogeneity estimates (standard deviations of participant-level random effects for M5 and M6) suppressed for brevity and reported in Web Appendix §F.4.

that adding these controls does not substantially shift the anchoring coefficient (0.039 vs. 0.038). This stability reflects a key feature of BRIDGE's architecture: because the similarity measures are computed from representations that are already partitioned to isolate attribute signal from nuisance variation, the nuisance controls are largely redundant. In contrast, when similarity is computed from unpartitioned full embeddings—which conflate focal attributes with non-focal variation (e.g., writing style)—the nuisance controls become essential for detecting the effect (see Benchmark Comparisons in Web Appendix §F.5). Moreover, as the Monte Carlo simulations demonstrate, BRIDGE's nuisance controls provide meaningful corrections when confounding variations are present. The coffee certification experiments further validated that the nuisance controls can provide additional safeguard, when a confound was present in the experiments by

design. Thus, BRIDGE presents a doubly robust approach to inference—on the one hand, its attribute representations are orthogonal to the nuisance variation and enable the specification of lower-dimensional and tractable measurement strategies; on the other hand, it develops nuisance controls for additional adjustment.

Robustness Check: Perturbation Analysis Results

We applied the perturbation analysis technique described previously. OpenAI's GPT-4.1 was prompted to slightly expand each description while maintaining tone, style, and core attributes (e.g., region, varietal). This process created a parallel dataset where descriptions were subtly elaborated upon without altering their fundamental meaning or factual content, allowing us to test BRIDGE's resilience to nuanced textual variations.

We conducted two key evaluations to confirm representations and key findings. First, we compared the numerical attribute representations derived by BRIDGE from the original and perturbed wine descriptions using the RV coefficient—a multivariate generalization of the squared Pearson correlation coefficient. Second, we reconstructed the variables for the anchoring analysis using the perturbed embeddings and re-estimated Models 1-4.

The RV coefficient analysis confirmed the robustness of BRIDGE's extracted representations to these perturbations. The RV coefficients comparing the original and perturbed representations were 0.972 for region representations and 0.980 for varietal representations, indicating a high degree of alignment. To assess whether these observed similarities were statistically significant, we conducted permutation tests comparing the observed RV coefficients to those expected under a null hypothesis of no systematic alignment. The resulting p-values were extremely low (all ps < 0.001 for both region and varietal representations), indicating that the alignment between original and perturbed representations was highly unlikely to be due to chance. These findings suggest that BRIDGE effectively treated the perturbations as nuisance variables, maintaining stable representations of the focal constructs of region and varietal.

The re-estimated statistical models using the perturbed embeddings also yielded results consis-

Table 4: Results: Perturbation Analyses of Wine Study

Effect	Model 1	Model 2	Model 3	Model 4
Intercept	0.082	0.083	0.083	0.084
	(0.011)	(0.012)	(0.011)	(0.012)
Similarity to First Task	0.036	0.041	0.036	0.042
	(0.013)	(0.014)	(0.013)	(0.013)
Similarity to Previous Task	0.012	0.017	0.012	0.018
	(0.013)	(0.014)	(0.013)	(0.014)
Similarity to Next Task	-0.004	-0.007	-0.005	-0.008
	(0.013)	(0.014)	(0.013)	(0.014)
Similarity to Final Task	-0.008	-0.008	-0.009	-0.008
	(0.013)	(0.014)	(0.013)	(0.014)
Similarity to Second Task	—	0.000	—	0.001
	—	(0.014)	—	(0.014)
Similarity to Two Tasks Before	—	-0.014	—	-0.013
	—	(0.014)	—	(0.014)
Similarity to Two Tasks After	—	-0.006	—	-0.006
	—	(0.014)	—	(0.014)
Similarity to Second-to-Final Task	—	0.018	—	0.018
	—	(0.014)	—	(0.014)
Nuisance Controls Included	No	No	Yes	Yes

Note. Fixed-effects estimates (posterior means). Standard errors reported in parentheses. Dashes (—) indicate the variable is not included. Exact posterior p-values and 95% credible intervals for all coefficients are reported in Web Appendix §F.4. Nuisance control fixed-effect estimates suppressed for brevity.

tent with those obtained from the original data. As shown in Table 4, the similarity to the first task (i.e., anchoring effects) remained significant across all models, with estimates ranging from 0.036 to 0.042 (ps range from 0.001 to 0.006). Other similarity measures, such as similarity to recent or subsequent tasks, remained nonsignificant, consistent with the original findings.

Benchmark Comparisons

To assess whether BRIDGE's neural architecture provides meaningful gains over simpler alternatives, we estimated the Model 3 specification using similarity measures derived from three alternative representation methods: raw embedding similarity, nuisance-only similarity, and linear projection. All benchmarks used the same underlying embeddings (OpenAI text-embedding-3-

large, 3,072 dimensions) and included BRIDGE-derived nuisance controls.

Raw Embedding Similarity. Computing cosine similarity directly in the full 3,072-dimensional embedding space yields a significant anchoring coefficient when nuisance controls are included ($\hat{\beta} = 0.215$, 95% CI [0.030, 0.398]). However, without nuisance controls, the same full-embedding similarity measure fails to reach significance ($\hat{\beta} = 0.177$, 95% CI [-0.005, 0.354]). This contrast is meaningful: recall that adding nuisance controls to the BRIDGE-derived similarity measures barely shifts the anchoring estimate (0.039 vs. 0.038). The difference is that BRIDGE's representations isolate attribute signal by construction, making nuisance controls redundant. When similarity is instead computed from unpartitioned embeddings that conflate attributes with stylistic variation, the nuisance controls become essential for recovering the effect. In both cases, the credible intervals remain substantially wider than BRIDGE's (0.368 and 0.359 vs. 0.059), reflecting the noise introduced by irrelevant dimensions.

Nuisance-Only Similarity. Computing similarity using only the nuisance dimensions—which capture description style, tone, and length orthogonal to the focal attributes—yields a null result ($\hat{\beta} = -0.007$, 95% CI [-0.030, 0.015]). This confirms that the anchoring effect is driven by attribute similarity, not by superficial textual resemblance between descriptions.

Linear Projection. As an alternative to BRIDGE's neural architecture, we estimated a linear projection baseline using Ridge regression to predict attributes from embeddings, followed by SVD on the residuals to construct nuisance controls. This approach produced a marginal result ($\hat{\beta} = 0.135$, 95% CI [-0.000, 0.269], $p = 0.051$), with a credible interval approximately 5 times wider than BRIDGE's. While the linear projection recovers the correct sign, its substantially wider credible interval reflects greater noise in separating attribute signal from nuisance variation. This empirical pattern aligns with the Monte Carlo simulations: with 426 regions and 708 varietals, the high-cardinality attribute space and the rich, diverse language of wine descriptions create precisely the conditions under which BRIDGE's nonlinear architecture provides substantive gains over linear decomposition.

Replacing BRIDGE's nuisance controls with 10 SVD-reduced Empath features does not alter

the anchoring estimate (0.040 vs. 0.039), confirming that the effect is robust to the choice of control specification; full benchmark results, including Empath psycholinguistic controls and attribute-only (Jaccard) similarity comparisons, are detailed in Web Appendix §F.5.

Validation: Replication with Controlled Coffee Stimuli and Conventional Analytical Methods

A study using controlled stimuli and conventional analytical methods replicated the observed product-anchoring effect in a different product category (coffee), providing converging evidence that the effect is not an artifact of BRIDGE. In this study, similarity was defined by attribute overlaps rather than AI-derived representations. Participants again exhibited preferences aligned with the attributes of options presented in their initial task. Full details of the experimental design, variable operationalization, model specifications, and results are provided in Web Appendix §G.

Study Discussion

We examined the dynamics of consumer preferences in sequential decision making for complex, verbally described products. Our findings support the notion that preferences are initially malleable but become defined by early (though arbitrary) product exposures, and remain largely invariant in later choices—thus exhibiting consistency. The results rule out revealed-preferences, fatigue, and preference-updating accounts. Perturbation analyses verify that our findings are robust to irrelevant modifications of the stimuli. A follow-up coffee experiment replicated the effect using a conventional, controlled design, establishing the effect as robust and generalizable across product categories and analytical approaches.

GENERAL DISCUSSION

We introduce an experimental design that accommodates many diverse real-world verbal stimuli while exerting statistical control for stimulus variability. At its core is BRIDGE: an interpretable AI model that transforms unstructured verbal stimuli into structured numerical representations of (1) focal variables and (2) orthogonal statistical controls for non-focal variation, facilitating

the analysis of participant responses to unstructured verbal descriptions for theory testing. This approach conceptually draws from recent evidence showing that LLM semantic encodings closely mirror human brain's operation for meaning and can be used to understand downstream tasks (Goldstein et al. 2025). BRIDGE facilitates stimulus sampling, strengthens a study's internal and external validity, and accommodates various theory testing paradigms including within-subject, between-subject, and mixed designs.

We conducted Monte Carlo simulations to investigate BRIDGE's effectiveness in establishing statistical control in experiments with many textual descriptions as stimuli. Even when the nuisance variation directly shifted preferences (main effect confounding) and moderated the strength of focal attributes (interaction effect confounding)—the two channels by which nuisance variation can confound experimental measurement—BRIDGE recovered the full set of structural parameters: treatment effects, interaction effects, and style-specific attribute effects. It matched the infeasible Oracle estimator, which is efficient (and hence ideal) but impracticable as it relies on information that is typically unavailable to the researcher (e.g., full confound structure). In contrast, both baseline models with no controls and traditional controls-only approaches, such as incorporating Empath lexical variables or deriving controls through linear projection, provided biased inference. Moreover, BRIDGE presents a feasible inference approach as it develops both nuisance controls and learned attribute representations using only information that is observed and available to the researcher.

In coffee certification experiments (with human participants) where a stimulus confound was deliberately introduced, BRIDGE—operating without explicit knowledge of this confound—recovered treatment effect estimates comparable to a stimulus-controlled, unconfounded baseline, even when a standard covariate adjustment was inadequate. These experiments show that BRIDGE can also be applied to conventional experimental designs with a small number of carefully constructed stimuli, as it provides additional safeguard for remaining nuisance variations.

To demonstrate BRIDGE with many diverse real-world stimuli, we studied preference dynamics in wine, where 1,000 consumers each evaluated 32 pairs randomly drawn from nearly 120,000

real-world tasting notes. Across participant choices for approximately 50,000 unique wines, the results revealed that preferences for verbally described products were initially malleable. Despite being entirely incidental, wines presented at the outset became product anchors that shaped later decisions. A controlled anchoring study replicated the effect. The results support the notion that preference dynamics are affected by the selective accessibility of knowledge about key attributes (e.g., region and varietal) as shaped by product anchors.

Contributions

We contribute to consumer behavior research in four ways. First, we present an alternative experimental design for studying consumer behavior in information-rich environments where products are conveyed in prose. Common contexts include hedonic consumption, complex financial products, and sustainability initiatives. For example, hedonic experiences are often described in prose to help consumers anticipate them. However, as Alba and Williams (2013) observe, "consumer researchers have been inclined to frame the issue narrowly, in part because many integral characteristics of hedonic consumption can be devilishly difficult to investigate via traditional experimental paradigms" (p. 3). Through BRIDGE, such contexts become experimentally tractable. In addition, BRIDGE can be used to resolve conflicting theoretical predictions arising from discrepancies in product presentations between simpler, stylized stimuli and detailed, real-world stimuli. BRIDGE can be used in both large-scale designs with unaltered real-world descriptions (as exemplified by our wine study) and conventional designs with controlled stimuli (as exemplified by our coffee certification experiments).

Second, BRIDGE addresses the stimulus-sampling problem. To mitigate this problem, scholars recommend enlarging the stimulus sample (Westfall, Judd, and Kenny 2015) and treating it as a random factor, viewing each stimulus as one of many possible items sampled from a larger population (e.g., Baribault et al. 2018; Judd, Westfall, and Kenny 2012). BRIDGE facilitates these design recommendations. While traditional experiments rely on orthogonal, factorial designs for causal inference, BRIDGE combines large-scale randomization, continuous attribute

representation, and statistical nuisance controls to minimize the likelihood that the observed effects are driven by extraneous features specific to a particular stimulus—a common concern in conventional experiments that rely on a small set of carefully pretested stimuli (Pham 2013; Simonsohn, Montealegre, and Evangelidis 2024)—improving internal validity. Furthermore, when stimuli are chosen randomly and presented in their original form, the observed effects are more likely to generalize beyond the sampled stimuli, improving external validity. Moreover, existing methods of manual coding are often impractical at scale. In our wine experiment, for instance, manually coding the relationships between randomly sampled wines would require computing approximately 2.5 billion distances. BRIDGE efficiently computes these distances, accounting for nuisance variables. Finally, presenting each participant with a distinct, randomly sampled subset of stimuli leverages the between-item variance highlighted by Judd, Westfall, and Kenny (2012), thereby increasing statistical power without increasing the participant sample size.

Third, our proposed research design enhances research objectivity and reproducibility. Traditional designs can involve numerous researcher decisions regarding (1) experimental design, such as stimulus simplification and selection, and (2) data analysis. These researcher degrees of freedom can inadvertently compromise research reproducibility. Our framework allows for the inclusion of real-world stimuli without abbreviation or stylization, broad sampling to reflect ecological distributions and achieve market coverage, and a structured, end-to-end methodological pipeline that can be pre-specified and pre-registered. These features can help mitigate the "garden of forking paths" problem (Gelman and Loken 2013), enhancing replicability and the overall credibility of the findings.

Fourth, BRIDGE builds on cognitive psychology and neuroscience research demonstrating that semantic embeddings align closely (and linearly) with the human brain's own hierarchical codes for speech and meaning (Goldstein et al. 2025), and can serve as a proxy for consumer conceptual knowledge in downstream tasks, such as food-health judgments (Gandhi et al. 2022). However, these embeddings are inherently unstructured and high-dimensional, comprising thousands of unlabeled dimensions, making it unclear which specific constructs or semantic features drive con-

sumer judgments. BRIDGE transforms semantic embeddings into structured, lower-dimensional, and interpretable representations—adaptable to specific contexts and dataset features through automatic tuning (see Web Appendix §F.2)—that can be used for theory testing. Our Monte Carlo simulations further establish that BRIDGE's nonlinear architecture is not merely a design choice but a methodological necessity; simpler linear decompositions of the same embeddings provide an incomplete account of the confounding factors, while BRIDGE recovers the full effect structure.

Limitations, Extensions, and Directions for Future Research

Expanding research contexts

Future research can use BRIDGE to study other text-rich consumer contexts. For example, researchers could study what makes online consumer reviews more useful by (a) collecting a corpus of reviews, (b) using BRIDGE to identify the underlying dimensions that affect review usefulness, and (c) validating the impact of these dimensions. To further verify these findings, researchers could develop controlled manipulations of the identified dimensions for use in a traditional laboratory study. This parallels our overall empirical strategy for assessing semantic anchoring with the BRIDGE-based wine experiment and its accompanying coffee validation study.

Consumer contexts often involve extensive descriptions of numerous aspects of an option. For example, insurance and financial products may be presented in rich, extensive texts, and consumers may consider multiple attributes simultaneously. Such contexts remain challenging to study using conventional methods. Given that BRIDGE can handle large, diverse product spaces (with high-cardinality categorical variables) through continuous representation, it offers a unique opportunity to examine how consumers navigate multiple dimensions in complex decision-making environments.

Our investigations focused on textual stimuli because they are prevalent and embedding-based measurement is currently more mature in text settings. However, as multimodal embedding models mature, BRIDGE can be extended to handle multimodal data inputs, including multimodal product descriptions. It would be interesting for future research to use BRIDGE to test consumer

behavior theories involving other forms of unstructured stimuli, such as images, audio, video, and even virtual and augmented reality.

An important direction is to assess the extent to which BRIDGE can be used to study abstract or holistic psychological constructs. Because BRIDGE uses supervised training to disentangle focal constructs from other variations in text, it is suited to contexts where the focal constructs can be reasonably defined and labeled. For instance, BRIDGE can be used to examine sentiment arcs in narratives (Toubia, Berger, and Eliashberg 2021) if these can be labeled in a training set. However, it is likely less suited for constructs where defining clear targets is difficult (e.g., assessing the overall 'persuasiveness' of myriad narrative structures without breaking them down into measurable components). Future research could integrate unsupervised topic modeling or clustering techniques prior to BRIDGE's supervised stage to help define dimensions within highly abstract texts.

Methodological extensions

A promising direction for future research is to use BRIDGE to inductively uncover latent dimensions from unstructured texts. To test preference dynamics in our wine study, we chose a context in which we know which attributes are likely important to consumers; this allowed us to deductively transform the unstructured tasting notes into a theory-based attribute representation. However, BRIDGE could be applied to inductively uncover latent dimensions. Future research could compare how theory-driven versus data-driven dimensions shape preference dynamics.

Researchers can also expand upon the stimulus sampling strategies employed in our studies. While large-scale random sampling ensures sufficient variation across the attribute space and approximates balance across attribute levels, typical sampling methods for participants (e.g., simple random, stratified, or cluster sampling) can be used for stimulus selection. Some sampling methods may be better suited for specific research questions, such as the use of stratified sampling (e.g., by wine provenance) to ensure balanced groupings when studying Old World versus New World wines. We discuss this flexibility in stimulus sampling methods and provide guidance in

the researcher's guide (Web Appendix §B).

Generative AI offers the ability to create descriptions when real-world stimuli are limited, such as in emerging or novel product categories[9], or when study objectives require highly controlled descriptions that mimic real products (e.g., the coffee descriptions in our certification and anchoring experiments). In these cases, researchers could adopt a human-in-the-loop framework, where humans guide the AI in generating stimuli or assist in selecting from AI-generated options (e.g., using platforms like MTurk). This approach enables researchers to craft realistic stimuli with ease and at a low cost, further broadening the applicability of BRIDGE-based experimental designs.

Managerial Implications

The use of real-world stimuli enhances the potential for counterfactual analysis, as the inferences drawn relate to actual products whose descriptions were used as stimuli, rather than to simplified and stylized abstractions. For example, our wine study estimates the preference structure of 1,000 consumers in the US, Australia, and New Zealand across a catalog of 119,955 wines. For wine producers and retailers, this detailed preference mapping could inform predictions of wine choices and preference-based metrics relevant to sales and market share under different product assortments or pairings. Much like structural econometric models that map economic theory to real-world behavior, these outputs may provide a foundation for marketing decision tools that inform the optimization of marketing strategies.

As AI advances, we hope BRIDGE enables researchers to design experiments with naturalistic stimuli that parallel how products are presented in the real-world marketplace, advancing the study of consumer behavior in information-rich environments.

[9]We leveraged this capability in our simulations and in our coffee certification studies, and illustrate it in our researcher's guide (see Web Appendix §B).

REFERENCES

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama (2019), "Optuna: A next-generation hyperparameter optimization framework," in *The 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–31.

Alba, Joseph W and Elanor F Williams (2013), "Pleasure principles: A review of research on hedonic consumption," *Journal of consumer psychology*, 23 (1), 2–18.

Amir, On and Jonathan Levav (2008), "Choice construction versus preference construction: The instability of preferences learned in context," *Journal of Marketing Research*, 45 (2), 145–58.

Ariely, Dan, George Loewenstein, and Drazen Prelec (2003), "'Coherent arbitrariness': Stable demand curves without stable preferences," *The Quarterly journal of economics*, 118 (1), 73–106.

Baribault, Beth, Chris Donkin, Daniel R Little, Jennifer S Trueblood, Zita Oravecz, Don Van Ravenzwaaij, Corey N White, Paul De Boeck, and Joachim Vandekerckhove (2018), "Metastudies for robust tests of theory," *Proceedings of the National Academy of Sciences*, 115 (11), 2607–12.

Bishop, Christopher M (1995), *Neural networks for pattern recognition*, Oxford university press.

Blackwell, Matthew and Michael P Olson (2022), "Reducing model misspecification and bias in the estimation of interactions," *Political Analysis*, 30 (4), 495–514.

Calder, Bobby J, Lynn W Phillips, and Alice M Tybout (1981), "Designing research for application," *Journal of consumer research*, 8 (2), 197–207.

Camerer, Colin (1997), *Rules for experimenting in psychology and economics, and why they differ*, Springer.

Caruana, Rich (1997), "Multitask learning," *Machine learning*, 28, 41–75.

Carvalho, Diogo V, Eduardo M Pereira, and Jaime S Cardoso (2019), "Machine learning interpretability: A survey on methods and metrics," *Electronics*, 8 (8), 832.

Chapman, Gretchen B and Eric J Johnson (1999), "Anchoring, activation, and the construction of values," *Organizational behavior and human decision processes*, 79 (2), 115–53.

Chernev, Alexander (2011), "Semantic anchoring in sequential evaluations of vices and virtues," *Journal of Consumer Research*, 37 (5), 761–74.

Clark, Herbert H (1973), "The language-as-fixed-effect fallacy: A critique of language statistics in psychological research," *Journal of verbal learning and verbal behavior*, 12 (4), 335–59.

Cook, Thomas D and Donald T Campbell (1979), *Quasi-experimentation*, Chicago, IL: Rand McNally.

Donkers, Bas, Benedict GC Dellaert, Rory M Waisman, and Gerald Häubl (2020), "Preference dynamics in sequential consumer choice with defaults," *Journal of Marketing Research*, 57 (6), 1096–1112.

Ebbesen, Ebbe B and Vladimir J Konecni (1980), "On the external validity of decision-making research: What do we know about decisions in the real world," *Cognitive processes in choice and decision behavior*, 21–45.

Einhorn, Hillel J and Robin M Hogarth (1981), "Behavioral decision theory: Processes of judgement and choice," *Annual review of psychology*, 32 (1981), 53–88.

Fast, Ethan, Binbin Chen, and Michael S. Bernstein (2016), "Empath: Understanding topic signals in large-scale text," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, ACM, 4647–57.

Frederick, Shane, Leonard Lee, and Ernest Baskin (2014), "The limits of attraction," *Journal of Marketing Research*, 51 (4), 487–507.

Gandhi, Natasha, Wanling Zou, Caroline Meyer, Sudeep Bhatia, and Lukasz Walasek (2022), "Computational methods for predicting and understanding food judgment," *Psychological Science*, 33 (4), 579–94.

Gelman, Andrew and Eric Loken (2013), "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time," *Department of Statistics, Columbia University*, 348 (1-17), 3.

Gigerenzer, Gerd (1991), "How to make cognitive illusions disappear: Beyond 'heuristics and biases'," *European review of social psychology*, 2 (1), 83–115.

Goldstein, Ariel, Haocheng Wang, Leonard Niekerken, Mariano Schain, Zaid Zada, Bobbi Aubrey, Tom Sheffer, Samuel A Nastase, Harshvardhan Gazula, Aditi Singh, and others (2025), "A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations," *Nature Human Behaviour*, 1–15.

Hoeffler, Steve and Dan Ariely (1999), "Constructing stable preferences: A look into dimensions of experience and their impact on preference stability," *Journal of consumer psychology*, 8 (2), 113–39.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989), "Multilayer feedforward networks are universal approximators," *Neural networks*, 2 (5), 359366.

Johnson, William B (1984), "Extensions of lipshitz mapping into hilbert space," in *Conference modern analysis and probability, 1984*, 189–206.

Judd, Charles M, Jacob Westfall, and David A Kenny (2012), "Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem." *Journal of personality and social psychology*, 103 (1), 54.

Kirk, R. E. (2013), *Experimental design: Procedures for the behavioral sciences*, SAGE Publications.

Kusupati, Aditya, Gantavya Bhatt, Aniket Rber, Matthew Wallingford, Aditya Sber, Prateek Jain, Sham Kakade, and Prateek Jain (2022), "Matryoshka representation learning," in *Advances in neural information processing systems*, 30233–49.

Latour, Kathryn A and Michael S Latour (2010), "Bridging aficionados' perceptual and conceptual knowledge to enhance how they learn from experience," *Journal of Consumer Research*, 37 (4), 688–97.

Levesque, Hector J (1986), "Knowledge representation and reasoning," *Annual review of computer science*, 1 (1), 255–87.

MacNeil, Karen (2015), *The wine bible*, Workman Publishing.

Martin, Leonard L, John J Seta, and Rick A Crelia (1990), "Assimilation and contrast as a function of people's willingness and ability to expend effort in forming an impression." *Journal of personality and social psychology*, 59 (1), 27.

Morales, Andrea C, On Amir, and Leonard Lee (2017), "Keeping it real in experimental research—understanding when, where, and how to enhance realism and measure consumer behavior," *Journal of Consumer Research*, 44 (2), 465–76.

Mussweiler, Thomas (2003), "Comparison processes in social judgment: Mechanisms and consequences." *Psychological review*, 110 (3), 472.

Payne, John W, James R Bettman, David A Schkade, Norbert Schwarz, and Robin Gregory (2000), "Measuring constructed preferences: Towards a building code," *Elicitation of preferences*, 243–75.

Pham, Michel Tuan (2013), "The seven sins of consumer psychology," *Journal of consumer psychology*, Elsevier.
Rabin, Matthew (1998), "Psychology and economics," *Journal of economic literature*, 36 (1), 11–46.
Rocklage, Matthew D, Derek D Rucker, and Loran F Nordgren (2021), "Emotionally numb: Expertise dulls consumer experience," *Journal of Consumer Research*, 48 (3), 355–73.
Rosenthal, Robert (1979), "The file drawer problem and tolerance for null results." *Psychological bulletin*, 86 (3), 638.
Sharot, Tali, Cristina M Velasquez, and Raymond J Dolan (2010), "Do decisions shape preference? Evidence from blind choice," *Psychological science*, 21 (9), 1231–35.
Simonsohn, U, A Montealegre, and I Evangelidis (2024), "Stimulus sampling reimagined: Designing experiments with mix-and-match, analyzing results with stimulus plots (SSRN scholarly paper 4716832)."
Slovic, Paul (1995), "The construction of preference." *American Psychologist*, 50 (5), 364.
Spicer, Jake, Jian-Qiao Zhu, Nick Chater, and Adam N Sanborn (2022), "Perceptual and cognitive judgments show both anchoring and repulsion," *Psychological Science*, 33 (9), 1395–1407.
Stone, Dan N and David A Schkade (1991), "Numeric and linguistic information representation in multiattribute choice," *Organizational Behavior and Human Decision Processes*, 49 (1), 42–59.
Strack, Fritz and Thomas Mussweiler (1997), "Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility." *Journal of personality and social psychology*, 73 (3), 437.
Tenenbaum, Joshua B, Charles Kemp, Thomas L Griffiths, and Noah D Goodman (2011), "How to grow a mind: Statistics, structure, and abstraction," *science*, 331 (6022), 1279–85.
Toubia, Olivier, Jonah Berger, and Jehoshua Eliashberg (2021), "How quantifying the shape of stories predicts their success," *Proceedings of the National Academy of Sciences*, 118 (26), e2011695118.
Wells, Gary L and Paul D Windschitl (1999), "Stimulus sampling and social psychological experimentation," *Personality and Social Psychology Bulletin*, 25 (9), 1115–25.
Westfall, Jacob, Charles M Judd, and David A Kenny (2015), "Replicating studies in which samples of participants respond to samples of stimuli," *Perspectives on Psychological Science*, 10 (3), 390–99.
Wickens, Thomas D and Geoffrey Keppel (1983), "On the choice of design and of test statistic in the analysis of experiments with sampled materials," *Journal of Verbal Learning and Verbal Behavior*, 22 (3), 296–309.
Wilson, Timothy D, Elliot Aronson, and Kevin Carlsmith (2010), "The art of laboratory experimentation," *Handbook of social psychology*, 1, 51–81.