Bridging the Gap: Using Interpretable AI to Incorporate Real-World Product Descriptions in Consumer Research Experiments

Anirban Mukherjee
Hannah Hanwen Chang
Sachin Gupta

15 June, 2025

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

Conventional experimental designs often require the use of simplified and stylized stimuli—constraints that may limit the realism, generalizability, and practical relevance of findings. This paper presents an AI-driven research design for consumer experiments that enables the use of myriad unaltered real-world product descriptions as stimuli. The authors develop an interpretable AI model termed labGPT to analyze experimental data. Comprising a partitioned deep learning neural network paired with a foundational large language model, labGPT generates structured, low-dimensional, and interpretable numerical representations that can be employed in statistical models of consumer responses, enabling hypothesis testing. The authors conduct Monte Carlo simulations showing that labGPT recovers true parameters even when textual stimuli included unobserved nuisance variations. They empirically study preference dynamics in a choice experiment with 1,000 consumers who were shown about 50,000 wine tasting notes randomly sampled from almost 120,000 wines in the market. The findings reveal that initial wines (despite being entirely incidental) shape subsequent preferences in the context of richly detailed verbal stimuli. By enabling the integration of complex, unstructured stimuli into experimental designs, the proposed approach enhances the realism, generalizability, and practical relevance of consumer research in complex information environments. Detailed researcher's guide and Python code are provided.

Keywords: Research Design, Interpretable Artificial Intelligence, Large Language Models, Consumer Behavior, Marketing Research, Marketing.

Real-world product descriptions can be complex, elaborate, and diverse. For instance, when choosing a vacation destination, consumers may encounter detailed verbal descriptions of destinations. 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. Other destinations may showcase entirely different experiences, with each description spotlighting different scenery, cultures, and activities. That inherent richness helps consumers anticipate these experiences, but poses a challenge for researchers: how much of this complexity should be incorporated into an experiment?

To study what drives consumer choices, experimental researchers tend to adopt one of two approaches. 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).

Alternatively, some researchers select one (or very few) real-world product descriptions (e.g., movie synopsis in Chang and Pham (2018), Study 5) as naturalistic stimuli that closely mimic the complex and diverse materials presented to consumers in the real world. The rich stimuli can help activate the same psychological processes that affect real-world consumer choices and enhance mundane realism[1] (Camerer 1997; Morales, Amir, and Lee 2017; Wilson, Aronson, and Carlsmith 2010), but also introduce unintended variability.

The two approaches reflect the inherent trade-off between control and generalizability in experimental design—an ongoing debate in consumer research framed as rigor versus relevance

[1]The extent to which the research setting is similar to the real world (Wilson, Aronson, and Carlsmith 2010).

(e.g., Lutz 2018; Lynch Jr et al. 2012; Pham 2013) or internal versus external validity (e.g., Calder, Phillips, and Tybout 1982; Lynch Jr 1982; Mukhopadhyay, Raghubir, and Wheeler 2018). Moreover, these design choices affect construct validity—whether the operationalized variables accurately reflect the intended theoretical constructs (Campbell and Cook 1979; Wells and Windschitl 1999).

Crucially, because conventional approaches tend to include only one or a handful of stimuli (compared to real-world product availability), they are subject to the stimulus-sampling problem (albeit to differing degrees): observed effects may be due to idiosyncrasies of the specific stimulus rather than the intended conceptual construct, so that the findings may not be generalizable to other stimuli (e.g., Clark 1973; Pham 2013; Wells and Windschitl 1999). This raises two concerns. First, it can undermine the study's internal validity when the stimulus sample for each experimental condition is biased with uncontrolled confounds such that confounds may be driving the effect, inflating Type 1 error (Judd, Westfall, and Kenny 2012). Moreover, Wells and Windschitl (1999) point out that "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' (Campbell and Cook 1979, p. 59)" (p. 1116). Second, it can challenge the study's external validity, because the use of a small sample of stimuli may not generalize to other stimuli in the larger category (e.g., Baribault et al. 2018; Judd, Westfall, and Kenny 2012). Scholars have recently speculated that many failures to replicate experimental results may stem from the stimulus-sampling problem (e.g., Westfall, Judd, and Kenny 2015).

In this research, we propose an experimental design that addresses the trade-off between experimental control (through simplified stimuli) and generalizability (through real-world stimuli). We are interested in contexts where (a) for each product, its verbal description is rich and (b) across products, their descriptions are varied and may not align with one another. These conditions amplify the stimulus-sampling problem yet are common in many consumer contexts, such as vacation planning, wine selection, and other hedonic experiences. Traditional experimental designs expose participants to a few select samples identical within conditions, allowing for conventional methods like ANOVA to compare participant responses across conditions. In contrast, our approach

allows each participant to be exposed to distinct stimuli randomly selected from a corpus of real-world product descriptions–to facilitate stimulus sampling–resembling a series of micro-experiments. This requires a novel analytical approach. The primary methodological challenge is the analysis of participants' responses to diverse, unstructured stimuli while maintaining control across varying stimulus characteristics.

To this end, we introduce labGPT, an interpretable AI model designed to process unstructured product texts. It employs a large language model (LLM) to generate high-dimensional representations of the unstructured texts, which it then transforms into structured, lower-dimensional, and interpretable representations. Each component of the output is designed to be a "knowledge representation" of a key focal attribute, serving as a numerical proxy for human cognition of that attribute that is amenable to computational reasoning (Carvalho, Pereira, and Cardoso 2019; Levesque 1986; Tenenbaum et al. 2011). This interpretability, combined with dimensionality reduction, ensures that the model's outputs maintain substantive meaning and are practical for use in statistical models of participant behavior. It also enables the development of statistical controls for nuisance variables, ensuring stimulus comparability in data analysis. Consequently, labGPT allows researchers to integrate large, diverse real-world product texts into experimental designs, enabling both within- and between-subjects hypothesis testing.

To demonstrate the effectiveness of labGPT, we conducted Monte Carlo simulations and empirical investigations using real-world stimuli and actual participant data. Our simulations provide evidence in controlled data conditions with known ground truth, showing that labGPT recovers true parameters even when textual stimuli included unobserved nuisance variations. We conducted an empirical study of preference dynamics in wine decisions, where 1,000 consumers evaluated 32 pairs randomly drawn from nearly 120,000 wines and described using real-world tasting notes. Across participant choices for about 50,000 unique wines[2], we find that consumer preferences are initially malleable. Once preferences are shaped by initially presented wines

[2]Each wine description averaged 53 words (SD = 11.86). Representative examples illustrating the complexity and variability of these wine descriptions are provided in Table A6 in Web Appendix §F.1. The inclusion of such rich and diverse stimuli would be impracticable using traditional experimental designs with the same number of participants.

(entirely incidental), subsequent decisions appear aligned with those first options (we term this "product anchors"). These results align with the notion that preference dynamics is influenced by the selective accessibility of knowledge regarding wine regions and varietals embedded in the unstructured texts. To further verify the phenomenon and ensure that it is not due to labGPT's analytic method, we conducted a follow-up experiment in a different category (coffee) using a conventional approach with controlled stimuli. The convergence of results across these studies—one analyzed using labGPT and the other using traditional methodology—establishes both the efficacy of labGPT as an alternative experimental design and supports our broader empirical findings. Taken together, the evidence shows that while complex, unstructured stimuli remain challenging for traditional methods, our analytical framework overcomes this hurdle and yields generalizable insights by statistically accounting for stimulus variability.

labGPT offers a promising alternative for investigating consumer behavior in contexts where product descriptions are rich and complex in the real world—and often conveyed in unstructured texts. Examples include hedonic experiences, financial products like mutual funds, diverse sustainability initiatives, and many others. Our approach allows an extensive array of real-world product descriptions to be used as experimental stimuli—a large random sample from the ecological distribution (range of products available in the market). It enables flexibility in several facets of experimental design: (1) improved stimulus sampling through the inclusion of a wider variety of real-world product descriptions, (2) increased statistical power through multiple distinct product presentations per participant, and (3) improved generalizability through greater overall coverage of product offerings in the market. While existing methods are valuable for theory testing using simpler stimuli, labGPT provides a complementary tool, allowing experiments to incorporate stimulus complexity akin to those in real-world settings. By bridging the gap between conventional experiments using controlled stimuli and the richness of real-world product descriptions, we hope our method can help researchers study consumer behavior in information-rich environments.

We organize our paper as follows. First, we introduce our proposed methodology, the data structures it generates, and the statistical models required for data analysis. We discuss why

existing analytical methods are not suitable for these data structures and outline the key properties necessary for theory testing and analysis. Next, we develop an interpretable AI model, labGPT, to address these properties, focusing on accuracy and tractability. These properties are crucial for method adoption by researchers and practitioners. We then present Monte Carlo simulations to assess labGPT's performance and demonstrate its application in an empirical investigation into a novel consumer behavior question that may be challenging to test using conventional experiments. We further report validation analyses, including both perturbation analyses and supplemental evidence from a follow-up experiment. Finally, we conclude with a discussion of the implications, limitations, and potential applications of our proposed methodology.

To complement this paper, we provide three key resources (with detailed Python code and extensive documentation as a Jupyter notebook): (1) a comprehensive researcher's guide (Web Appendix §B) detailing the analytical model (transformation of unstructured stimuli into interpretable representations and their subsequent use in theory testing); (2) perturbation analyses (Web Appendix §C) to assess and validate the performance of labGPT; and (3) detailed numerical simulations (Web Appendix §D) to demonstrate the efficacy and robustness of the proposed methodology, facilitating research accessibility and transparency. Through these resources, we aim to make our methodology more accessible to a wider audience, for researchers interested in using our methodology or adapting it to a different context and study.

METHOD DEVELOPMENT

Traditional experimental designs in consumer research typically involve a few standardized conditions, where participants within each condition are exposed to identical stimuli. This standardization enables researchers to apply conventional statistical methods to compare responses across conditions. For example, in a typical study, participants' responses can be modeled using a linear model:

$$y_i = \alpha + \beta D_c(i) + \gamma s(i) + \epsilon_i, \tag{1}$$

where y_i represents the response of participant i; α is the intercept; β is the hypothesized effect and γ are coefficients for covariates describing systematic factors. The independent variables are $D_c(i)$, a dummy variable indicating whether participant i is in condition c, and $s(i)$, any covariates (e.g., participant characteristics). 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, γ, and $s(i)$), in addition to accommodating interaction terms between the condition dummies and the covariates of interest.

It is typical for stimuli to be identical for participants in the same condition. In this scenario (and if $s(i) = 0$), hypothesis testing on β is equivalent to performing ANOVA, assessing whether group means differ significantly by analyzing variances within and between groups. If $s(i) \neq 0$, hypothesis testing on β is equivalent to performing ANCOVA, which similarly assesses whether group means differ significantly by analyzing variances within and between groups while accounting for covariates $s(i)$. However, this analytical approach assumes that the stimuli are comparable across conditions and any observed differences in responses can be attributed solely to the experimental manipulation $D_c(i)$ or covariates $s(i)$. When participants are presented with diverse, unstructured real-world stimuli—such as complex product descriptions—the stimuli may vary in numerous ways and any extraneous variations embedded within them, known as nuisance variables, can confound the effects of the focal variables. Prior studies have shown that traditional designs, which treat stimuli as fixed (and fail to account for variations across stimuli), can inflate the empirical Type 1 error rate (Judd, Westfall, and Kenny 2012; Wickens and Keppel 1983).

One strategy might be to manually code the nuisance variables, translating them into control variables. However, manually coding nuisance variables may be unfeasible due to the many nuanced differences (e.g., tone, style, voice, formality) that are hard to identify and isolate (Clark 1973), but may influence respondent behavior. 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. While this characteristic may be irrelevant to the research question, it may still influence consumer behavior. In addition, this

approach may be unfeasible due to scale. For instance, our empirical study employed 50,000 distinct stimuli, selected from a set of 120,000 descriptions—a substantial sample size that would likely rule out manual coding.

We propose an alternative strategy: the use of AI to systematically extract statistical controls for the nuisance variables and incorporate them into the analysis. Specifically, we introduce an interpretable AI model, labGPT, to transform the complex and unstructured stimuli into interpretable, low-dimensional, and structured numerical representations. This transformation allows us to specify statistical models that account for stimulus variability and facilitate robust hypothesis testing and theory development. In particular, we consider linear models of the form:

$$y_i = \alpha + \beta D_c(i) + \gamma s(i) + \delta c(i) + \epsilon_i, \tag{2}$$

where δ represents coefficients for the statistical controls of nuisance variables, and $c(i)$ are the controls, which are algorithmically extracted from the stimuli. $c(i)$ are distinct from $s(i)$ in that, whereas $c(i)$ are nuisance variable controls derived from labGPT, $s(i)$ are any observed differences; the latter of which can be controlled in traditional experiments.

To illustrate our approach, consider controlling for the stimuli's color in an experimental setting. One strategy might be to present participants with stimuli of random colors and include color as a covariate in the model (i.e., as $s(i, j)$, for the j^{th} stimulus). By randomizing color and incorporating this information as a covariate, we can rule out color as a confounding factor when estimating β. Now, suppose that the color of each stimulus is unknown to the researcher *a priori*. A possible strategy might be to use an algorithm to automatically code a control variable indicating color, enabling the aforementioned approach. This is analogous to the approach we propose.

We provide a complementary approach for experimental design and analysis. In contexts where real-world verbal descriptions are sufficiently simple, traditional methods may suffice. However, in more complex contexts where such descriptions are rich, existing experimental design methods may impose substantial constraints on the types of research questions that can be

investigated and the extent to which the study stimuli approximate real-world descriptions. In such scenarios, we view our proposed methodology as a complement to existing approaches.

We organize this section as follows. First, we introduce labGPT, our custom AI system for extracting interpretable representations and nuisance controls from unstructured stimuli. We discuss the conceptual background for using natural language processing and AI in this context, linking knowledge representations from cognitive psychology with LLMs and AI systems. We provide a detailed explanation of how labGPT operates. Next, we outline the general step-by-step procedure for conducting a study using our proposed research design and introduce perturbation analysis as a key technique for validating the robustness of the methodology. To facilitate adoption, the web appendix provides a detailed researcher's guide to our model, including Python code for all analytical components (see Web Appendix §B). To conclude this section, we discuss how labGPT's structured, algorithmic approach enhances research objectivity and reproducibility, facilitating research practices like pre-registration and stimulus sampling.

labGPT: An Interpretable AI Model for Theory Testing

At the core of human cognition lies knowledge representation. Humans represent information about the external world and the stimuli they encounter, using this information for decision-making and problem-solving (Markman 2013). Similarly, LLMs that emulate human cognition encode unstructured inputs into high-dimensional numerical knowledge representations—embeddings in a vector space. This is because the encoding process is essential for enabling LLMs to understand and utilize human knowledge; when systematically encoding, processing, and retrieving information, LLMs mirror the intricate relationships between human concepts and how humans organize and access knowledge (Mikolov et al. 2013). Embeddings capture human conceptual understanding while remaining amenable to mathematical analysis (Brown et al. 2020; Radford et al. 2019).

Converging evidence supports this theorizing between LLM encodings and human knowledge representations. LLMs reliably predict human cognitive processes, including brain activity (Goldstein et al. 2025; Storrs et al. 2021), vocabulary comprehension (Griffiths, Steyvers, and Tenenbaum

2007; Landauer and Dumais 1997), and lexical semantic categorization (Landauer, Foltz, and Laham 1998). For example, Goldstein et al. (2025) show that encoding models based on OpenAI's embeddings linearly predict neural activity along the brain's speech-language pathway–from hearing sounds to understanding what is said–during everyday conversations. Other studies apply LLMs to investigate consumer evaluations and behaviors, such as similarity and healthfulness judgments (Battleday, Peterson, and Griffiths 2021; Gandhi et al. 2022; Peterson, Abbott, and Griffiths 2018). Taken together, a growing body of literature in cognitive psychology and neuroscience show how LLM encodings can approximate human conceptual knowledge for downstream tasks (e.g., Bhatia and Richie 2022; Goldstein et al. 2025; Laverghetta Jr et al. 2022).

Nevertheless, three key challenges emerge when considering the use of LLM encodings for hypothesis testing. LLM encodings are inherently unstructured—they lack explicit organization. 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. And they lack interpretability; that is, 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.

To address these challenges, we propose an interpretable AI model (labGPT) that combines an LLM with a two-stage specialized neural network architecture to develop structured, low-dimensional, and interpretable attribute representations from verbal descriptions. First, the LLM is used to develop LLM encodings. Second, a fully connected feedforward network compresses the original high-dimensional LLM encodings into a lower-dimensional intermediate representation (Johnson 1984). This network, illustrated in Figure 1 as taking a 12-dimensional LLM Encoding (labeled Input Layer in the figure) and processing it through two sequential 8-node hidden layers (labeled Hidden Layer 1 and Hidden Layer 2), produces an intermediate representation.[3]

[3]Figure 1 depicts the labGPT architecture as applied to a wine description (an exemplar of the complex and expressive data the model is designed to handle). The figure uses arbitrary (and few) nodes and layers for expositional clarity. In practice, labGPT is designed to adapt to context (depending on dataset features, such as textual intricacy and attribute granularity) through automatic tuning (e.g., Keras Tuner). For instance, our empirical application uses

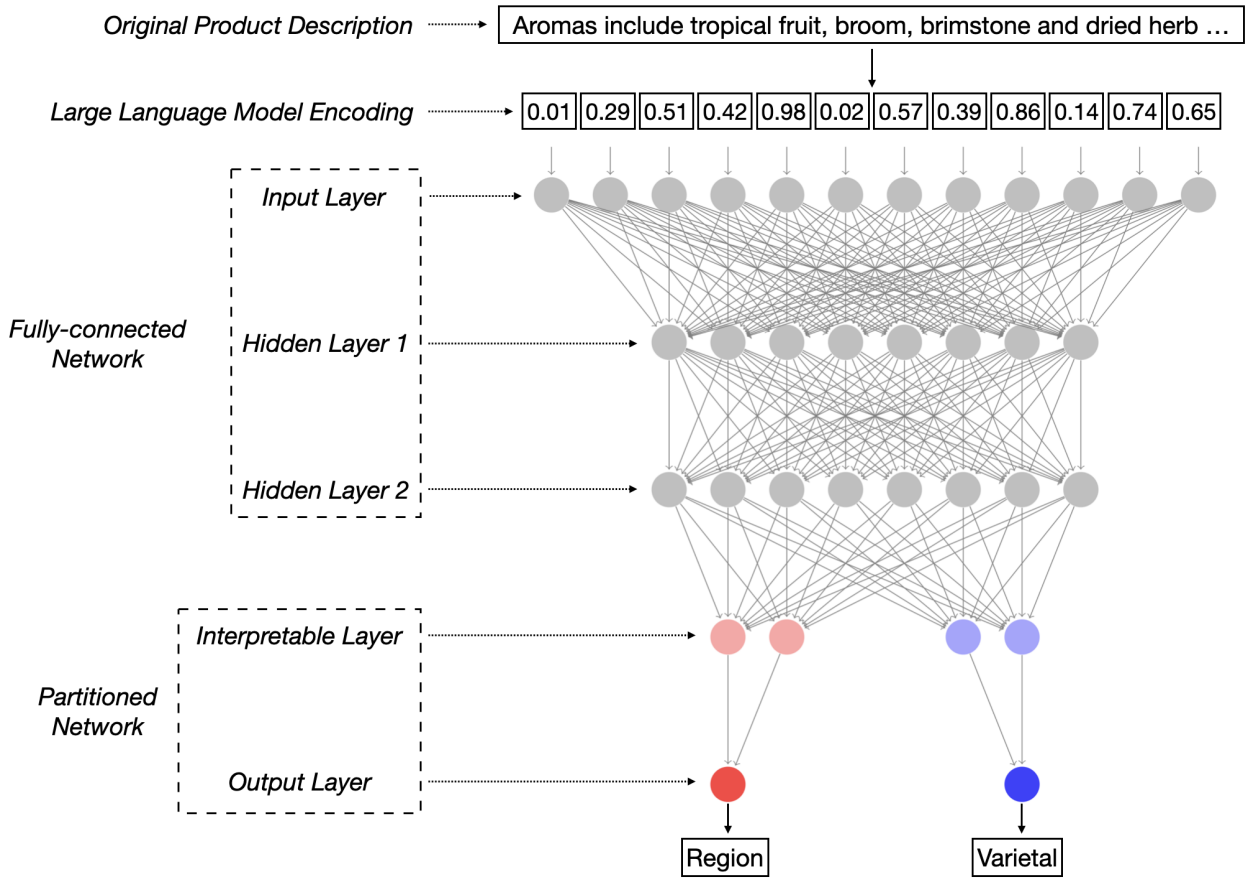


Figure 1: Schematic of labGPT Architecture

Note: A simplified illustration of the labGPT architecture. An LLM maps an unstructured, real-world product description to an encoding. A fully connected network (a 12-dimensional input layer and two 8-node hidden layers) compresses the LLM encoding to produce a 8-dimensional intermediate representation. A partitioned network (a 4-node interpretable layer, partitioned into 2 attribute-specific components) maps the intermediate representation to 2 distinct attributes. The components of the partitioned network are depicted in blue and red.

A partitioned feedforward network then maps the intermediate representation (i.e., the output of Hidden Layer 2 in the figure) to an interpretable layer that is partitioned into distinct components. Each component is dedicated to predicting a distinct attribute during training.[4] Consequently, labGPT learns to compress the LLM encoding and extract attribute-relevant information (Bishop 1995; Hornik, Stinchcombe, and White 1989), embedding it in an attribute-specific vector space (i.e., the vector space of an attribute's representations). The output from each component of the

OpenAI's text-embedding-3-large, which generates a 3,072-dimensional encoding. The final interpretable layer in our study has 16 nodes (corresponding to 16 output dimensions, 8 for each of 2 attributes). This adaptability in configuration is a key strength of labGPT.

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

interpretable layer serves as the corresponding attribute's representation.

Thus, labGPT learns to perform both data compression (via its first network) and the extraction of attribute-specific information (via its second network), as is required for the effective analysis of participant responses to unstructured stimuli. To further ensure interpretability in the attribute-specific components, the (external) LLM encoding is frozen during training, and labGPT is trained using the following multi-term loss function.

The primary term of the loss function focuses on accurately predicting the attributes (e.g., province, varietal) associated with each output node. In addition, the function includes a contrastive loss term that seeks to encourage the attribute-specific vector spaces to be more distinct (i.e., it encourages the representations of one attribute to be distinct from the representations of other attributes), and another contrastive loss term to encourage the representations of the distinct levels within each attribute to be distinct from one another (i.e., for each attribute-specific vector space to be expansive). The loss function balances interpretability and performance, as the primary loss 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 loss terms benefit from larger dimensional spaces, as these facilitate orthogonality in the attribute- and attribute-level-specific representations.

Finally, statistical controls for the nuisance variables are developed by projecting the LLM encodings 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 our proposed research design involves five key steps, depicted in Figure 2 and elaborated upon in the researcher's guide (see Web Appendix §B). The procedure,

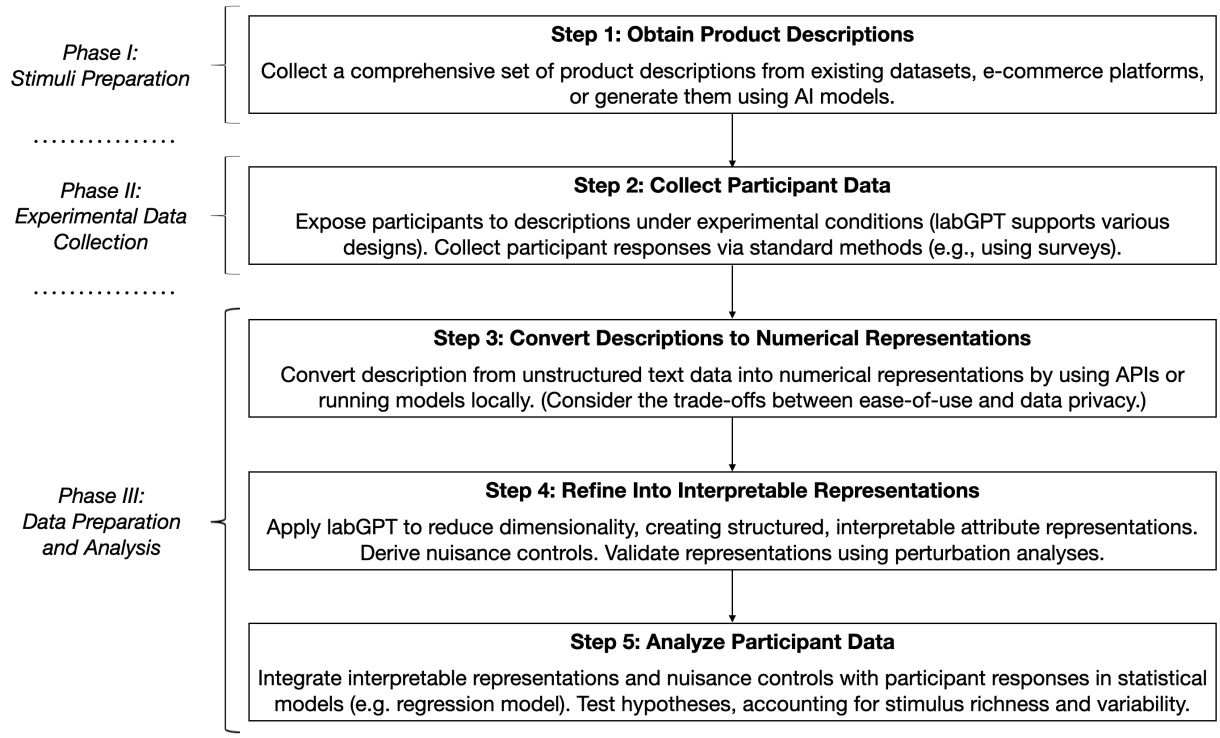


Figure 2: Procedure of a Study Using labGPT

supported by detailed Python code provided in the web appendix, is modular and adaptable, enabling researchers to modify key elements—such as stimuli, experimental manipulations, or statistical models—without extensive reconfiguration. The steps offer a structured yet flexible workflow that accommodates a wide range of experimental contexts and research objectives. Below, we provide an overview of each step, along with references to the corresponding sections in the researcher's guide for further details.

1. **Obtain Product Descriptions**: Begin by collecting a comprehensive set of product descriptions. These can be sourced from existing datasets, collected from e-commerce platforms, or generated using AI models. Ethical considerations, such as compliance with privacy regulations (e.g., GDPR) and data-sharing agreements, must be addressed throughout this process. *For detailed guidance on sourcing, generating, and ethically handling product descriptions, refer to Step 1 in the researcher's guide (Web Appendix §B.1.1).*

2. **Collect Participant Data**: Expose participants to the compiled product descriptions under

experimental conditions. Our design supports various experimental formats, including within-subjects, between-subjects, and mixed designs. Collect participant responses using standard methods such as surveys or choice tasks, adhering to established best practices in experimental research. *The researcher's guide (Web Appendix §B.1.2) provides further details on integrating these complex stimuli into standard data collection procedures.*

3. **Convert Descriptions to Numerical Representations**: Transform the unstructured product descriptions into high-dimensional numerical representations ("embeddings") using a LLM. This encoding process captures the semantic richness of the descriptions. Researchers can choose between API-based models (e.g., OpenAI's text-embedding-3-large, used in our wine study) and locally hosted, open-source alternatives (e.g., Nvidia's NV-Embed-V2). API-based models can be highly cost-effective; for instance, embedding approximately 120,000 unique descriptions in our wine study incurred minimal costs (around $2). Alternatively, locally hosted models offer stability, complete budget control, and enhanced data privacy. *Step 3 of the researcher's guide (Web Appendix §B.1.3) provides a comprehensive walkthrough of this process, including code examples and considerations for selecting an appropriate LLM.*

4. **Refine into Interpretable Representations**: Refine the high-dimensional embeddings using labGPT, a partitioned neural network designed to generate structured, interpretable, and low-dimensional representations. labGPT maps each attribute of interest to a distinct numerical component, facilitating clear comparisons across diverse stimuli.[5] Additionally, labGPT derives statistical controls for nuisance variables by orthogonalizing intermediate representations, ensuring extraneous textual variations do not confound the analysis. Validate these representations using perturbation analyses to confirm robustness against irrelevant textual variations. *The researcher's guide (Web Appendix §B.1.4) provides detailed instructions and code for implementing labGPT, including model training, validation, and the*

[5]In the spirit of transfer learning common in GPT models, the representations derived from labGPT for specific attributes (such as region and varietal in our wine study) can be viewed as specialized, pre-trained embeddings for this domain. Researchers working with the same stimuli could potentially leverage these existing representations directly for related analyses, bypassing the need for retraining. We thank the review team for this insight.

generation of statistical controls.

5. **Analyze Participant Data**: Integrate the refined representations with participant responses to perform rigorous data analysis. By combining attribute-specific encodings and nuisance controls in regression models or other statistical frameworks, researchers can test hypotheses about consumer behavior while accounting for the richness and variability of the stimuli. *Step 5 of the researcher's guide (Web Appendix §B.1.5) discusses how to integrate these representations into statistical models, including code examples and guidance on interpreting the results.*

Validating labGPT Robustness: Perturbation Analyses

A potential concern with AI-driven analyses, such as those using labGPT, is their sensitivity to superficial changes in input text that are irrelevant to the core constructs being studied. To address this concern and validate the robustness of our research findings, we present *perturbation analyses*. This method systematically evaluates whether labGPT's extracted representations—and the resulting conclusions drawn from them—remain stable when controlled, irrelevant modifications are introduced to the textual stimuli. The objective is to ensure that the focal constructs identified by the model (e.g., attributes) are not unduly influenced by these extraneous textual changes, and to verify that the method's nuisance controls effectively capture such incidental variability.

The core principle involves introducing controlled modifications (perturbations) designed to alter stylistic elements without changing core attributes, thereby simulating realistic textual noise. Using methods detailed in Web Appendix §C, such perturbations can include appending irrelevant marketing phrases (e.g., 'Limited stock!') or employing an LLM to subtly adjust tone or phrasing. These perturbed descriptions are presented to the trained model to generate perturbed variants of the attribute representations.

The evaluation comprises two steps. First, we assess the stability of the numerical representations derived from original versus perturbed text. This can involve calculating similarity metrics (like the RV coefficient) between the original attribute representations and the perturbed attribute representations. In this case, a high similarity indicates robustness. Alternatively, a

lower similarity between the original nuisance controls and the perturbed nuisance controls indicates successful capture of the perturbations. Second, we repeat the final statistical analysis (e.g., regression models) using the perturbed representations and compare the resulting coefficient estimates, significance levels, and overall model fit to those obtained using the original representations. Stability across these evaluations confirms that observed effects stem from genuine (focal) theoretical constructs rather than superficial (non-focal) textual artifacts. Full implementation details, code examples, and further discussion of perturbation analyses as a validation tool are provided in Web Appendix §C.

Enhancing Objectivity and Pre-Registration in Experimental Research

Crucially, our proposed research design enhances research objectivity and reproducibility.[6] Traditional experimental designs involving unstructured stimuli often require numerous subjective decisions about stimulus selection, simplification, and analysis—decisions that can introduce researcher degrees of freedom and unintentionally compromise reproducibility. Our framework transforms this process into a structured, *end-to-end algorithmic pipeline* where key components can be precisely pre-specified and pre-registered before data collection or analysis. For instance, researchers can commit *a priori* to: an *embedding method* (e.g., a specific versioned API or open-source model like NV-Embed-v2), the *labGPT architecture* or a *systematic procedure* for determining it (e.g., using automated tools like Keras Tuner within defined parameters, as demonstrated in our Web Appendix §F.2), a *stimulus sampling strategy* (e.g., random sampling from a real-world corpus) to reduce stimulus-selection bias, and an *analytical plan*, including independent and dependent variables, derived representations, nuisance controls, and models for hypothesis testing.

By leveraging real-world stimuli and automating the generation of representations and controls within this pre-defined algorithmic framework, the entire process can be executed automatically. This approach addresses the "garden of forking paths" problem (Gelman and Loken 2013), where seemingly innocuous analytical decisions can drastically alter research conclusions. By reducing

[6]We thank the review team for pointing this out.

subjectivity and minimizing researcher degrees of freedom related to stimulus selection, data processing, and data analysis, this approach strongly supports open science practices like pre-registration, enhancing replicability, and bolstering the overall credibility of research findings by limiting opportunities for (1) biased stimulus selection (Wells and Windschitl 1999) and (2) post-hoc analytical decisions (Gelman and Loken 2013). Our detailed researcher's guide and accompanying code (Web Appendix §B) are designed to facilitate such transparent and reproducible workflows.

MONTE CARLO STUDIES EVALUATING LABGPT'S PERFORMANCE

Before applying labGPT to real-world experimental data, we validate its performance under controlled conditions. The study design, Python code, results, and extensive discussion are presented in Web Appendix §D. Below, we provide a synopsis.

We aim to evaluate labGPT's ability to recover the true underlying parameters of a data-generating process for experimental participant data in the presence of unobserved variations (nuisance variables) in the stimuli. We examine two challenges: (1) isolating focal variables from extraneous nuisance variables (e.g., stylistic variations) embedded within the product descriptions, and (2) effectively incorporating these nuisance variables into econometric models of participant responses to ensure consistent and precise inference. By creating synthetic data with known ground truth, we can assess how accurately labGPT isolates focal attributes, controls for nuisance variation, and captures experimental effects—even when product descriptions vary systematically in unobserved variables.

We center our simulations on the following hypothetical marketing research scenario: investigating whether presenting information about an environmental issue in one domain (operationalized as overfishing) can activate related concerns and influence preferences in another (specifically, for organic versus non-organic packaged salads). We test labGPT's ability to infer this spillover effect despite confounding nuisance variables (unobserved stylistic variations); our focus here is on methodological validation, not the substantive question itself.

To simulate key challenges in the use of real-world stimuli, we systematically vary the style

of product descriptions across three types: Factual, Engaging, and Creative. Specifically, in the simulation, participants are presented with product descriptions that vary in attributes (organic, size, type, weight) and are randomly assigned to either a control group or a treatment group. To simulate an unobserved correlation between nuisance variables and the focal variables of interest, the treatment group is exposed to a higher proportion of Engaging and Creative descriptions (8 Engaging, 12 Creative, and 4 Factual descriptions per participant across 24 tasks), while the control group sees a higher proportion of Factual descriptions (12 Factual, 8 Engaging, and 4 Creative descriptions per participant across 24 tasks)—we treat the differences in styles as unobserved and therefore potential confounds.

To evaluate labGPT's effectiveness in accounting for unobserved stylistic differences, we compare its performance against several analytical approaches, including three key benchmark models (summarized in Table 1):

1. **Traditional (No Controls):** A traditional model that omits nuisance controls entirely. This reflects a common approach when such variations are unobserved but is likely to yield biased estimates in our confounded setup.
2. **Traditional (Empath Controls):** A feasible benchmark representing conventional text analysis approaches. It uses general-purpose text features derived from the open-source Empath library, mirroring the practice of using summary psycholinguistic variables (like those from LIWC). Since LIWC is proprietary, we construct an open-source equivalent by applying Singular Value Decomposition to the approximately 200 category scores generated by Empath for each description. The resulting low-dimensional Empath features serve as general-purpose nuisance controls, offering reproducibility and accessibility but potentially lacking the specificity to fully capture targeted confounds.
3. **Oracle (Style Dummies):** An infeasible 'oracle' model that directly incorporates dummy variables for the unknown description styles. This represents the theoretical best possible performance under perfect information regarding the nuisance variable, but is unrealistic in practice where style is unobserved.

4. **labGPT:** Our proposed model, which uses interpretable AI to extract low-dimensional, attribute-specific representations and orthogonal nuisance controls from the high-dimensional LLM embeddings. This approach is designed to flexibly and automatically control for unobserved confounds in unstructured text.

Table 1: Comparison of Analytical Approaches

Model	Controls for Style?	How?	Feasible?
No Controls	No	None	Yes
Empath Controls	Partial	Empath features	Yes
Oracle (Style Dummies)	Yes	Style dummies	No
labGPT	Yes	LabGPT nuisance controls	Yes

Programming Framework and Statistical Models

We implement the comparison using a simulation involving the following eight steps:

1. **Setup, Ground Truth, and Base Data:** Establish the true causal parameters for experimental effects (e.g., treatment effect, attribute preferences, interaction) and generate a base set of product descriptions systematically varying the focal attributes (in this case: organic status, size, type, and weight of salads).[7]
2. **LLM-Based Product Description Generation:** Use a LLM to create multiple stylistic variations (e.g., Factual, Engaging, Creative) for each base description. This introduces realistic textual diversity and serves as the *unobserved stylistic confound* central to the simulation's challenge.
3. **Embed the Product Descriptions:** Convert all generated textual descriptions into high-dimensional numerical embeddings using a standard embedding model (OpenAI's text-embedding-3-large).

[7]Specific attribute levels used were: Size (small, medium, large, extra-large), Type (spring mix, iceberg lettuce, radicchio, spinach, arugula), Weight (5–20 oz).

4. **Generate Empath Controls:** Create conventional text-based controls to serve as a feasible benchmark. This involves applying a standard lexical analysis library (Empath) to the descriptions and reducing the resulting feature vectors (via Singular Value Decomposition).
5. **Train the Interpretive AI Model (labGPT):** Train the labGPT neural network on the embeddings (from Step 3) to produce two key sets of predictors: (1) Low-dimensional, interpretable representations corresponding to the focal product attributes. Nuisance controls derived from the residual embedding variation *orthogonal* to the attribute representations.
6. **Simulate Participant Responses:** Generate synthetic participant responses (e.g., ratings) using a utility function. This function incorporates the ground truth parameters (Step 1), the product's focal attributes, the description's style (Step 2), the experimental condition, and random error. The simulation design ensures that assignment to the experimental condition is correlated with the style of description encountered, thereby introducing the intended omitted variable bias challenge.
7. **Apply labGPT and Evaluate Performance:** Estimate regression models predicting the simulated participant responses (Step 6) using the interpretable attribute representations and orthogonal nuisance controls (from Step 5).
8. **Estimate Benchmark Models and Compare Performance:** Estimate regression models using different sets of predictors:
    - *Empath:* Attribute predictors plus the benchmark lexical controls (from Step 4).
    - *Oracle:* Attribute predictors plus dummy variables indicating the true (but typically unobserved) description style (feasible only in simulation).
    - *No Controls:* Attribute predictors only, ignoring style.

Product j is characterized by the tuple $\left(\text{Organic}_j, \text{Size}_j, \text{Type}_j, \text{Weight}_j, \text{Treatment}_j, \text{Style}_j\right)$. Organic_j is a binary indicator for whether the product is organic; Size_j and Type_j are categorical variables representing the product's size and type, respectively; Weight_j is a continuous variable for the product's weight; Treatment_j is a binary indicator for experimental condition (treatment or control); and Style_j represents the style of the product description (Factual, Engaging, or Creative),

with the distribution of styles varying systematically between the treatment and control groups. Participant i's latent utility for product j is modeled as follows:

$$\begin{aligned}\text{preference}_{ij} = \alpha + \beta_1 \text{Treatment}_j + \beta_2 \text{Organic}_j + \beta_3 (\text{Treatment}_j \times \text{Organic}_j) \\ + \beta_4 \text{Size}_j + \beta_5 \text{Type}_j + \beta_6 \text{Weight}_j + \beta_7 \text{Style}_j + \varepsilon_{ij},\end{aligned}$$

where α is the intercept; β_1, β_2, and β_3 capture the treatment effect, the organic premium, and their interaction (spillover effect), respectively; β_4, β_5, and β_6 represent the effects of size, type, and weight, respectively; β_7 quantifies the effect of description style; and ε_{ij} is the random error.

The style effect term $\beta_7 \text{Style}_j$ acts as the unobserved confound. The *No Controls* model ignores this term entirely, risking omitted variable bias. The *Empath Controls* model attempts to capture this term using general-purpose text features. The *labGPT model* uses nuisance controls derived from the orthogonalized residuals of the embeddings after accounting for focal attributes. The *Oracle model* requires infeasible knowledge: it directly includes style dummies, and therefore perfect knowledge, to control for the unknown style confound. Comparing how well each feasible approach (No Controls, Empath Controls, labGPT) recovers the key causal parameters β_1 and β_3 in the presence of the confounding $\beta_7 \text{Style}_j$ term allows us to evaluate their practical effectiveness.

Results and Discussion

The Monte Carlo simulations highlight a critical challenge in experimental research: the presence of unobserved variables, such as stylistic variations in product descriptions, that can confound traditional analyses. Table 2 presents the coefficient estimates for the four models, illustrating this challenge and demonstrating the relative effectiveness of different approaches in addressing it: (1) our labGPT estimator, which incorporates nuisance controls derived algorithmically from the product descriptions; (2) a traditional estimator without explicit controls for style (i.e., without controls for nuisance variables); (3) a traditional estimator that includes style dummies (an infeasible 'oracle' model that has impracticable knowledge on the nature and distribution of nuisance variables); and (4) a traditional estimator using feasible controls derived from Empath features.

While feasible in this simulation, the oracle model (3) is impractical in real-world settings where variables such as style are commonly unobserved. The Empath features model (4) is, instead, the typical compromise in applied research.

The results reveal that the traditional estimator without style controls produces severely biased and inconsistent estimates for the key causal parameters. As shown in Table 2, the coefficient for the treatment effect (Condition, β_1) is significantly overestimated (0.887) compared to the ground truth (0.25), representing a relative bias of over 250%, and its 95% confidence interval [0.842, 0.933] excludes the true value. This bias arises because the unobserved stylistic variations (Factual, Engaging, Creative) are correlated with both the treatment assignment and the outcome, creating an omitted-variable problem. Specifically, participants in the treatment group are more likely to see Engaging and Creative descriptions, while those in the control group are more likely to see Factual descriptions. When style is omitted from the model, its effect (β_7) is erroneously attributed to the treatment (β_1), leading to inaccurate causal inference. This underscores a core issue with traditional approaches when dealing with unobserved confounds in complex, unstructured data.

In contrast, the traditional estimator that includes style dummies accurately recovers the true parameters, as shown by the close alignment of its estimates with the ground truth (e.g., Condition estimate of 0.220). This model serves as an important, albeit unfeasible, benchmark: it represents the best possible performance under perfect information about the confounding nuisance variables. Its reliance on observing and quantifying style, however, makes it impractical in real-world scenarios where such nuances are typically unobserved and costly to label.

The Empath controls model, representing a feasible and typical text analysis approach, reduces the bias compared to the naive model but fails to fully recover the ground truth. The estimated treatment effect is 0.466 (SE = 0.021, 95% CI: [0.424, 0.508]). While this is a substantial improvement over the naive estimate (0.887), it is still nearly double the true value (0.25), and the confidence interval does not include the true value. This suggests that while general-purpose lexical controls derived from libraries like Empath can mitigate some confounding, they lack the specificity to fully capture nuisance variables, resulting in continued bias in causal estimates. The model fit

Table 2: Comparison of Model Estimates from Monte Carlo Simulation

	labGPT	**No Style**	**Oracle**	**Empath**
Condition ($\beta_1 = 0.25$)	0.239 (0.021)	0.887 (0.023)	0.220 (0.020)	0.466 (0.021)
	[0.198, 0.280]	[0.842, 0.933]	[0.181, 0.260]	[0.424, 0.508]
	✓	✗	✓	✗
Condition × Organic ($\beta_3 = 0.50$)	0.496 (0.029)	0.517 (0.033)	0.521 (0.027)	0.489 (0.029)
	[0.438, 0.553]	[0.452, 0.581]	[0.469, 0.574]	[0.432, 0.546]
	✓	✓	✓	✓
Organic ($\beta_2 = 0.50$)	0.498 (0.020)	0.505 (0.023)	0.497 (0.019)	0.525 (0.021)
	[0.458, 0.537]	[0.460, 0.551]	[0.460, 0.534]	[0.485, 0.566]
	✓	✓	✓	✓
Weight ($\beta_6 = 0.10$)	0.101 (0.002)	0.100 (0.002)	0.099 (0.001)	0.100 (0.002)
	[0.098, 0.104]	[0.097, 0.104]	[0.097, 0.102]	[0.097, 0.103]
	✓	✓	✓	✓
R-squared	0.608	0.401	0.605	0.534
Controls Included	Nuisance	None	Dummies	Empath
labGPT Controls	Yes	No	No	No
Style Dummies	No	No	Yes	No
Empath Controls	No	No	No	Yes
Size Dummies	Yes	Yes	Yes	Yes
Type Dummies	Yes	Yes	Yes	Yes

Standard errors in parentheses, 95% confidence intervals in brackets.
✓ indicates true value falls within 95% CI, ✗ indicates it does not.
Ground truth values: Condition(β_1)=0.25, Condition×Organic(β_3)=0.50, Organic(β_2)=0.50, Weight(β_6)=0.10.
'labGPT' includes labGPT nuisance controls, 'No Style' includes no controls, 'Oracle' includes style dummies, and 'Empath' includes Empath controls. All models include size and type dummies.

($R^2 = 0.535$) also remains considerably lower than the Oracle and labGPT models.

By comparison, labGPT offers a practical and effective solution to this challenge. By algorithmically deriving nuisance controls that capture the unobserved stylistic variations, labGPT recovers all true underlying parameters reported in Table 2 with high accuracy and precision. The estimated treatment effect (0.239) closely aligns with the ground truth (0.25), and its confidence interval [0.198, 0.280] includes the true value. Similarly, the interaction effect estimate (0.496) is very close to the ground truth (0.50) and its confidence interval [0.438, 0.553] contains the true value. Furthermore, the main effects for Organic (0.498 vs. 0.50) and Weight (0.101 vs. 0.10) are accurately recovered. This demonstrates the method's effectiveness in mitigating the confounding

bias and facilitating accurate inference across the model. The overall model fit ($R^2 = 0.608$) is also comparable to the infeasible Oracle model ($R^2 = 0.605$).

Overall, these simulations demonstrate the varying effectiveness of different approaches for analyzing complex, unstructured data in experimental settings. Traditional methods omitting controls yield biased results for causal parameters in the presence of confounding nuisance variables. Feasible alternatives using general-purpose text features like Empath controls can partially mitigate this bias but may be insufficient, still resulting in biased causal inference. The infeasible oracle model sets an accuracy benchmark but is impractical in real-world settings. labGPT offers a feasible approach that achieves high accuracy in recovering causal effects–comparable to the oracle model–by generating tailored, orthogonal nuisance controls. This makes it an effective solution for rigorous causal inference in experiments with real-world verbal stimuli.

PREFERENCE DYNAMICS IN UNSTRUCTURED REAL-WORLD PRODUCT DESCRIPTIONS

Normative theories view consumer preferences as stable and merely retrieved during decision making (Rabin 1998). This view has been challenged by mounting evidence showing that preferences are often constructed based on how options are presented in the decision-making environment (Payne et al. 2000; Slovic 1995). This constructive process can extend to subsequent decisions as consumers learn their preferences, navigate attribute trade-offs, and build confidence in their judgments (Amir and Levav 2008; Hoeffler and Ariely 1999; Yoon and Simonson 2008).

Making repeated choices can stabilize preferences. These findings align with studies on 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 the anchor to semantic words (Chernev 2011). Previous studies on preference dynamics typically employ structured choice contexts in which options are defined by two simpler, aligned attributes (Amir and Levav 2008; Donkers et al. 2020; Yoon and Simonson 2008).

However, consumers in real-world settings often encounter options described with unstruc-

tured 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 either choose their preferred option (our wine study) or indicate a relative preference (our coffee study). For each participant, each product in the sequence is drawn randomly (without replacement) from about 120,000 wines or 36 coffees.

We posit that preferences for such 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 (Mussweiler and Strack 2001; 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.

Testing this hypothesis ideally involves presenting to participants a large, randomly selected set of real-world descriptions, such that each product is described in prose, initial products (the anchor) are incidental, and the characteristics across products may not align. labGPT can help facilitate this empirical examination. It can transform complex prose descriptions (unstructured

texts) into (1) structured, low-dimensional, and interpretable numerical representations of focal product attributes, which helps us compute similarity metrics, and (2) statistical (nuisance) controls for non-focal, confounding variations across the unstructured texts (e.g., stylistic features). Thus, labGPT facilitates the analysis of experiments using large, ecologically valid stimulus sets, which would be challenging via conventional experiments.

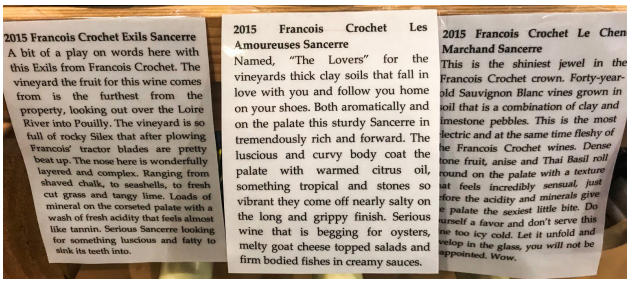
The purpose of our empirical investigation is twofold. First, we test our anchoring hypothesis in sequential decisions for products described in prose. Second, we demonstrate and validate labGPT as a methodology enabling such research. We begin with a large-scale experiment using real-world wine descriptions–an ideal context due to its reliance on complex, subjective language. We analyze participants' choices via the representations generated by labGPT. To provide converging evidence—and to ensure our results reflect the phenomenon itself rather than artifacts of any method—we conduct a follow-up experiment in a different category (coffee) using controlled stimuli amenable to traditional analysis techniques. Together, these studies allow us to gauge the generalizability of the anchoring effect and the validity of our proposed methodology.

Context and Product Descriptions

The domain of wine consumption presents a suitable context for our research question. Wine, valued for its sensory and hedonic experiences (Bisson et al. 2002), is a category characterized by substantial differentiation (Lynch Jr and Ariely 2000). Each wine's unique profile—shaped by its region of origin (terroir) and grape varietal (MacNeil 2015; Robinson 2015)—is typically conveyed in unstructured texts. This differs substantially from simplified stimuli often used in controlled experiments. 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. We next describe how we apply the labGPT approach to this study.

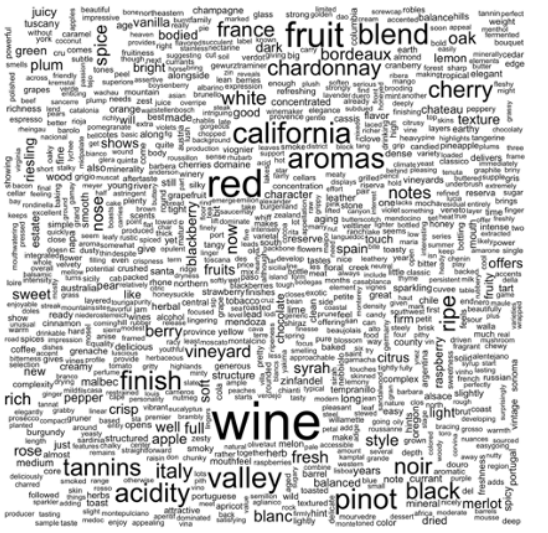
As an initial step (corresponding to Step 1 in Figure 2), we obtain a comprehensive dataset containing the verbal descriptions (tasting notes) of 119,955 wines from Wine Enthusiast, a globally recognized publication. These tasting notes are generated through extensive blind taste tests,

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. For instance, "shelf talkers" (small descriptive cards attached to store shelves) help consumers navigate their choices by providing descriptions of the wines' characteristics (see panels 3a and 3b in Figure 3 for examples). We incorporated this extensive set of real-world wine descriptions in our experiment, facilitated by our proposed methodology. This helps enhance the study's mundane realism by mirroring the materials consumers encounter in real-world retail environments. It also improves ecological validity—by sampling a large, diverse corpus of real-world product descriptions, the observed findings are more likely to generalize to other real-world wine choices.



(a) Close-up of shelf talkers.

(b) Shelf talkers in context.



(c) Word cloud of taste descriptors.

Figure 3: Illustrative Wine Descriptions and Vocabulary.

Notes: Panels (a) and (b) show wine shelf talkers. Panel (a) provides a close-up view, while Panel (b) shows them in a retail context (image from 'The pros and cons of shelf talkers,' by Vicki Denig at SevenFifty Daily, reproduced with permission). Panel (c) displays a word cloud of the 1,000 most frequent taste descriptors found in the dataset.

Figure 3c (see panel (c) in Figure 3), a word cloud of the top 1,000 taste descriptors in the tasting notes, illustrates the inherent complexity of the vocabulary used in the tasting notes; representative examples illustrating the complexity and variability of wine descriptions are provided in Web

Appendix §F.1 (Table A6). Specifically, the vocabulary used to describe wine tastes is remarkably rich and varied, ranging from common terms like "dry," "full-bodied," or "oaky" to more unusual descriptors such as "flinty," "barnyard," or "bruised." Occasionally, it might appear peculiar, capturing unique notes like "green bell pepper," "petrol," "wet stone," or "bubblegum." This extensive vocabulary poses a key challenge for traditional experimental methods that often rely on simpler stimuli, underscoring the need for labGPT.

Our conceptualization is that initially presented options activate selective semantic knowledge: participants draw on the more accessible, relevant attributes in the initial products and use them as reference points for later evaluations (Mussweiler 2003; Mussweiler and Strack 2001; Strack and Mussweiler 1997). In the context of wines, those diagnostic cues tend to be "region" and "varietal," which appear in tasting notes and why retailers 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). Moreover, varietal-region attributes are intrinsically linked to a wine's taste, its cultural roots, and geographical-climatic provenance (MacNeil 2015). As these attributes are crucial for consumers evaluating wines (Latour and Latour 2010; Rocklage, Rucker, and Nordgren 2021), participants are likely to access and draw on region and varietal when forming preferences. Focusing our analysis on these attributes therefore maps onto our proposed mechanism: early activation of salient varietal-region concepts in the initial products serve as the semantic reference frame that shapes subsequent decisions.

Our final dataset comprises descriptions of 119,955 wines from 427 wine-growing regions and 708 wine-grape varietals. Each record includes a wine's name, region, varietal, and tasting note (averaging 53 words in length; SD = 11.86 words). We use these descriptions to conduct a study on consumer preference dynamics in sequential decision-making.

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. Consequently, our findings are relevant to various businesses, such as wineries, wine distributors, and retailers.

Each participant completed 32 sequential decision 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 indicate their preferred option. Each wine was described by its name and tasting notes (including varietal and region), mirroring the information presented to consumers in retail settings. This demonstrates labGPT's flexibility, as it (a) allows participants to encounter many different options and (b) presents wines in prose descriptions, resembling how consumers encounter these products in the real-world.

We opted for the relatively long sequence of 32 decision tasks, even though our method accommodates any length (including a single task). Extending the sequence 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.

The stimulus-randomization procedure, where each wine is selected randomly without replacement from the product corpus, offers three methodological benefits. First, each participant encountered a unique initial product anchor (the first pair). As the anchor is arbitrary, no single wine (or its varietal or region) can drive the effect, reducing potential bias from a specific stimulus.

Second, randomization ensured exogenous variation in our key measure of semantic similarity. Given that participants encountered a unique sequence of options, the *similarity* between options in any given task and the options in the initial anchor task (or any other candidate anchor task) varied randomly across participants and sequences, providing a clean, participant-specific

treatment effect for estimating the anchoring effect.

Third, randomization allowed us to conduct placebo tests, where options presented at later sequence positions can be treated as pseudo-anchors to assess whether they (incorrectly) predict earlier choices. Since participants cannot be aware of options in future tasks, and the set and sequence of options was random, the similarity of options in the current task to those in future tasks served as placebo tests: hypothesis tests wherein the same analysis is applied to variables where no real effect is expected or possible, helping to rule out alternative explanations. For example, in our wine study, we analyze the relationship between current choices and similarity to future options, expecting a null result, as any significant finding would suggest our method is prone to identifying artifacts (spurious correlations) rather than true anchoring effects.

Overall, by leveraging the labGPT method, our study effectively addressed the stimulus-sampling problem identified in the literature (e.g., Baribault et al. 2018; Wells and Windschitl 1999) and other alternative explanations (to be discussed in the study's results section).

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. Participants reported strong interest (M = 5.59) and liking (M = 6.22) for wine on seven-point scales (1 = 'not at all'; 7 = 'very'). Their task involvement, measured using three items (e.g., 'I could relate to the overall situation of evaluating wines'; 1 = 'strongly disagree,' 7 = 'strongly agree'; $\alpha = .74$), was high (M = 5.73). Participants also found the wine descriptions moderately easy to understand (M = 5.43 on a 7-point scale from 'very difficult' to 'very easy'). All observations were included in our analyses (i.e., no data exclusions).

Variable Operationalization and Model

To test our anchoring hypothesis—that preferences revealed in subsequent choices align with initial product anchors—we developed measures based on the cosine similarity of the labGPT-

derived attribute representations of the options. Specifically, for each option presented in a given choice task (left or right), we first calculated its similarity to a specific *candidate anchor task*. This 'option-level similarity score' was defined as the sum of the cosine similarity between that option's attribute representations and the attribute representations of the two options shown in the anchor task.[8] Higher scores indicate greater similarity between a presented option and the options previously seen in the anchor task.

Our binary dependent variable was coded 1 if the left option was chosen and 0 if the right option was chosen. The key independent variable used to predict this choice was the *difference* between the option-level similarity scores for the left and right options relative to each specific anchor task. A positive coefficient for this differenced variable would indicate that participants were more likely to choose the option that was more similar to the options presented in the anchor task (an assimilation effect). A negative coefficient would indicate a preference for the more dissimilar option to those in the anchor task (a contrast effect). A non-significant coefficient would suggest that similarity to that anchor task did not explain participants' choices. Based on this operationalization, we defined the following variables by changing the position of the anchor in the sequence:

1. *Similarity to Options in the First Task (Sim^{First}):* For each option in the current task, we summed its similarity to each of the options in the first task. We then differenced the similarities of the two options in the current task. A positive and significant coefficient on this variable would provide support for our hypothesized anchoring effect.

2. *Similarity to Options in the Previous Task (Sim^{Prev}):* We followed the same procedure as for Sim^{First}, replacing the options in the first task with the options in the previous task. The coefficient on this variable tests for the recency effect.

[8]The specific labGPT architecture used to generate these representations was systematically 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 province 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.

3. *Similarity to Options in the Next Task (Sim^{Next}):* We followed the same procedure as for $\text{Sim}^{\text{First}}$, replacing the options in the first task with the options in the next (i.e., subsequent) task. As the options in the next task are unknown at the time of decision-making, the coefficient on this variable serves as a placebo test. We expect a nonsignificant coefficient, ruling out alternative explanations such as preference heterogeneity.

4. *Similarity to Options in the Last Task (Sim^{Last}):* We followed the same procedure as for $\text{Sim}^{\text{First}}$, replacing the options in the first task with the options in the last task. As the options in the last task are also unknown at the time of decision-making, the coefficient on this variable also serves as a placebo test.

For a closer examination of preference dynamics and to rule out alternative explanations, for each task t, we additionally calculated similarity to the second task ($\text{Sim}^{\text{Second}}$), two tasks before the current task ($\text{Sim}^{\text{Two Before}}$), two tasks after the current task ($\text{Sim}^{\text{Two After}}$), and the second-to-last task ($\text{Sim}^{\text{Second Last}}$) using the same procedure. Each variable was defined only when the corresponding anchor task fell within the valid task range (e.g., $\text{Sim}^{\text{Two Before}}$ was defined only for tasks $t \geq 3$, and $\text{Sim}^{\text{Two After}}$ was defined only for tasks $t \leq 31$).

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

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

ε_{ikt} is i.i.d. Gumbel, yielding a logit choice model.

To systematically test our hypothesis while ruling out potential alternative explanations, we estimate a sequence of nested fixed-effects models. *Model 1* includes only the four focal similarity terms ($\text{Sim}^{\text{First}}$, Sim^{Prev}, Sim^{Next}, Sim^{Last}) to test our hypothesis, with the recency (Sim^{Prev}) and

placebo (Sim^{Next}, Sim^{Last}) terms as internal controls. Building on this, *Model 2* incorporates the four additional similarity measures ($\text{Sim}^{\text{Second}}$, $\text{Sim}^{\text{Two Before}}$, $\text{Sim}^{\text{Two After}}$, $\text{Sim}^{\text{Second Last}}$) to further assess any observed anchoring effect. Separately, *Model 3* augments Model 1 by adding the five labGPT-derived nuisance controls (Nuisance_q) to account for irrelevant stylistic variations in the stimuli descriptions. *Model 4* then integrates all components, estimating a model with all eight similarity measures and the five nuisance controls included, representing the fully specified fixed-effects model.

To account for individual differences in sensitivity to these factors, we next introduce participant-level heterogeneity. *Model 5* builds upon the specification of Model 3, allowing the coefficients for the intercept, the four focal similarity effects ($\text{Sim}^{\text{First}}$, Sim^{Prev}, Sim^{Next}, Sim^{Last}), and the five nuisance control effects (δ_{qi}) to vary across participants. *Model 6* introduces participant-level heterogeneity based on the fully specified Model 4. This model allows *all* fixed-effect coefficients—intercept, all eight similarity effects, and all five nuisance control effects—to vary by participant, providing the most flexible representation of individual differences.

Results

The results, detailed in Table 3, indicate that participants anchored their preferences to the wines presented in the initial task, consistently favoring wines with similar attributes in subsequent choices.[9] The coefficient for "Similarity to First Task" is positive and credibly different from zero across all seven models with this predictor, including these four models (posterior mean ranging from 0.038 to 0.041, 95% CIs exclude 0), those that include only single predictors (see Web Appendix §F.4.1), and those that omit nuisance controls and heterogeneity (Models 1 and 2; see Web Appendix §F.4.2). In contrast, the coefficients for similarity to other candidate anchors (e.g., wines shown in the previous, next, or last tasks) and the additional similarity terms were non-significant across all of these specifications.

[9]Full results for single-predictor models are provided in Table A9 in Web Appendix §F.4.1, fixed-effects models (M1, M2, M3, and M4) including nuisance controls in Table A10 in Web Appendix §F.4.2, and the fully heterogenous models (M5 and M6) including all fixed and random effects in Table A11 in Web Appendix §F.4.3.

Table 3: Results: Wine Study

Effect	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Intercept	0.083***	0.083***	0.089***	0.090***
	(0.012)	(0.012)	(0.013)	(0.013)
Similarity to First Task	0.038***	0.039***	0.038**	0.041**
	(0.013)	(0.013)	(0.015)	(0.015)
Similarity to Previous Task	0.015	0.016	0.015	0.016
	(0.013)	(0.014)	(0.014)	(0.015)
Similarity to Next Task	-0.010	-0.009	-0.010	-0.009
	(0.013)	(0.014)	(0.014)	(0.014)
Similarity to Last Task	-0.005	-0.004	-0.004	-0.002
	(0.013)	(0.014)	(0.015)	(0.016)
Similarity to Second Task	—	-0.001	—	-0.003
	—	(0.014)	—	(0.016)
Similarity to Two Tasks Before	—	-0.011	—	-0.011
	—	(0.014)	—	(0.015)
Similarity to Two Tasks After	—	-0.008	—	-0.012
	—	(0.014)	—	(0.014)
Similarity to Second-to-Last Task	—	0.012	—	0.011
	—	(0.014)	—	(0.016)
LOOIC	38694	38700	38320	38292
Nuisance Controls Included	Yes	Yes	Yes	Yes
Full Heterogeneity Included?	No	No	Yes	Yes

Note. Fixed-effects estimates (posterior means) from Bayesian hierarchical logistic regressions. Standard errors are reported in parentheses. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Dashes (—) indicate that the variable is not included in the model. 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.4.3.

The nonsignificant coefficient for "Similarity to Previous Task" indicates that the data does not support recency accounts. Moreover, the coefficients for "Similarity to Next Task" and "Similarity to Last Task" act as placebo tests: as participants were unaware of the options in later tasks, their current choices should not be driven by similarity to these future options. Their nonsignificance aligns with this expectation and suggests the observed effect of initial (incidental) anchor is not an artifact of unobserved preference heterogeneity correlating choices across random tasks.

Model fit comparisons using the Leave-One-Out Information Criterion (LOOIC) reveal that

accounting for individual differences substantially improves explanatory power. Models 5 and 6, which incorporate participant-level heterogeneity, provide a considerably better fit than their fixed-effects counterparts (Models 3 and 4, respectively). Among the models presented, Model 6, which allows all fixed effects from Model 4 to vary across participants, achieves the best overall fit (LOOIC = 38292).[10] These findings suggest that once an initial reference point is established, consumers maintained coherent preferences aligned with the characteristics of those initial options, even after accounting for nuisance variables and substantial individual-level heterogeneity in preferences. Moreover, we 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 was non-significant, suggesting that the anchoring effect does not diminish over the experiment, ruling out fatigue as an alternative explanation.[11] These results are detailed in Web Appendix §F.4.4.

labGPT played a crucial role in developing our independent measures. Our conceptual framework posits that the initial options activate selective semantic knowledge about key product attributes (region and varietal), rather than superficial textual differences. To capture this, labGPT generated low-dimensional numerical representations for each attribute that were, by construction, orthogonal to nuisance variables. This orthogonalization ensured that the resulting similarity measures reflected only the similarity of these attributes, uncontaminated by potentially confounding factors such as description style—a feature critical to theory testing in our experimental design.

Robustness Check: Perturbation Analysis Results

To confirm the robustness of our findings, we applied the perturbation analysis technique described previously. Perturbations were generated using OpenAI's GPT-4.1, prompted specifically 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 labGPT's resilience to nuanced textual variations.

[10] Detailed LOOIC comparisons are provided in Web Appendix §F.4.
[11] We thank the review team for this suggestion.

We conducted two key evaluations to confirm representations and key findings. First, we compared the numerical attribute representations derived by labGPT 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 labGPT's extracted representations to these perturbations. The RV coefficients comparing the original and perturbed embeddings were 0.972 for region embeddings and 0.980 for varietal embeddings, 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 ($ps < 0.0001$ for both region and varietal embeddings), indicating that the alignment between original and perturbed embeddings was highly unlikely to be due to chance. These findings suggest that labGPT 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 consistent 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 (all $ps < 0.001$). Other similarity measures, such as similarity to recent or subsequent tasks, remained nonsignificant, consistent with the original findings.

Validation Study: Experiment Using Controlled Stimuli (Coffee Descriptions)

We conducted a validation study to further verify the anchoring phenomenon observed in our wine study and offer convergent evidence against method-specific artifacts. To provide evidence from a different product category, we situate our study in coffee which, like wine, offers rich sensory experiences and is often presented in texts. We employed conventional design with experimentally controlled descriptions and conventional analytical techniques. This allows us

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 Last 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-Last Task	—	0.018	—	0.018
	—	(0.014)	—	(0.014)
Nuisance Controls Included	No	No	Yes	Yes

Note. Fixed-effects estimates (posterior means) from Bayesian hierarchical logistic regressions. Standard errors are reported in parentheses. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Dashes (—) indicate that the variable is not included in the model. Nuisance control fixed-effect estimates suppressed for brevity.

to verify the findings observed in the wine study, generalize the anchoring phenomenon to a different category, and show that the findings reflect the phenomenon and not due to the method.

Specifically, we focused on two key coffee attributes—aroma and taste—each with six possible levels (aroma: floral, fruity, nutty, chocolatey, spicy, earthy; taste: bitter, sweet, acidic, sour, umami, balanced). This structure yielded 36 unique attribute combinations. To generate coffee profiles, we used ChatGPT's autocomplete feature to create a short, standardized description for every aroma–taste pairing. Specifically, we prompted ChatGPT with the stem: *"This coffee's [adjective] [aroma level] and [adjective] [taste level]…"* such that each description (1) referenced the correct attributes, (2) maintained a neutral and concise tone, and (3) followed a single-sentence structure. To make the mock texts appear as realistic coffee profiles, we prompted ChatGPT to include a

short phrase on the sensory experience of the aroma-taste pairing, while keeping the underlying aroma–taste attributes constant. The coffee stimuli averaged 24 words in length (median = 24 words), with comparable detail and complexity across the set. By standardizing and aligning these attributes, we could leverage the attribute overlaps as a similarity metric.

A total of 362 undergraduate students (69.9% women, average age = 21.6) participated in this study in exchange for course credit. Each participant evaluated 18 pairs of coffee descriptions, indicating their relative preference on a 7-point scale (1 = strongly prefer Option A, 7 = strongly prefer Option B). Each coffee in each pair was randomly selected, without replacement, from the set of 36 coffees. While all participants were presented with the same set of coffees, the specific pairings and sequences varied across participants. After completing the evaluation tasks, participants provided basic demographic information (gender, age). Detailed methods are available in Web Appendix §E.

To test the effect of initial product anchors on participant preferences, we defined a similarity measure based on the extent of the match between the attribute levels of the options in each task and those in the initial task. This served as our key independent measure of similarity, where a larger value indicated greater alignment with the options shown in the initial task.

To rule out alternative explanations, we implemented three controls. First, the options shown in the initial task, as well as all subsequent options, were chosen at random, ensuring that any observed similarity effects were not driven by systematic patterns in the order of the options. Second, we computed additional similarity measures to test for recency effects and placebo effects. Specifically, akin to the measures in our wine study, we calculated similarity to the options shown in the previous task (to test recency) and similarity to options in the next and last tasks (as placebo tests); we expected non-significance for these latter measures if primacy drove choices. Third, further analyses included similarity measures relative to other tasks (e.g., the second task, tasks two steps before/after, and the second-to-last task). The non-significance of these additional measures would suggest that these other positions in the sequence did not affect focal preferences.

Under these controlled conditions, we replicated the anchoring effects observed in our main

wine study. The results, detailed in Web Appendix §E, revealed that similarity to the options presented in the *first* task significantly predicted participants' subsequent preferences ($ps < 0.001$), while similarities to options from previous, next, last, or other temporally distant tasks did not (all NS). This overall pattern, combined with the randomization, makes unobserved heterogeneity an unlikely explanation and confirms that the anchoring effect observed in our main study is not contingent on wines nor an artifact of labGPT.

Study Discussion

We examined the dynamics of consumer preferences in sequential decision-making for complex, verbally described products, focusing on how initially encountered products influence subsequent choices. The findings revealed a consistent pattern of preference alignment, whereby participants generally favored options similar to initially presented options (though entirely incidental). Specifically, similarity to the wines encountered in the first task significantly predicted subsequent choices, whereas similarities to wines from previous, next, or final tasks did not.

The robustness of this anchoring effect was confirmed through two validation exercises. First, perturbation analyses verified that our core findings—particularly the significant influence of similarity to the initial task—remained stable even when irrelevant textual modifications were intentionally introduced into the product descriptions. This indicates that the anchoring effect was driven by underlying product attributes captured by labGPT rather than superficial linguistic or stylistic variations. It shows that labGPT could account for these extraneous textual modifications.

Second, a validation study replicated the anchoring phenomenon, using controlled stimuli from a different product category (coffee) and employing conventional analytical methods. In this study, similarity was defined by explicit attribute overlaps rather than AI-derived representations. Participants again exhibited preferences aligned with the attributes of coffee options presented in their initial task. Thus, this replication provided converging evidence supporting both our substantive conclusions regarding preference anchoring and the methodological efficacy of labGPT in generalizing the effect to real-world products.

Taken together, these findings support the notion that consumer preferences are initially malleable but are defined by early (though arbitrary) product exposures and become largely invariant in later choices, thereby exhibiting consistency. Our results rule out revealed-preferences, fatigue, and preference-updating accounts, reinforcing the idea that initial exposures to verbally described products can serve as powerful reference points that guide later choices in information-rich environments. The effect emerges as a robust and generalizable phenomenon, providing convergent validation of our proposed research design and methodology (labGPT).

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 labGPT: an explainable AI model that transforms the unstructured verbal descriptions into structured representations of (1) focal variables and (2) statistical controls for non-focal variations, facilitating analysis of participant responses to unstructured stimuli. It is conceptually grounded in recent evidence showing LLM encodings closely mirror human brain's operation for speech and meaning and can be used to understand downstream tasks. labGPT transforms the inherently unstructured LLM encodings through a two-stage network into structured, low-dimensional, interpretable representations that can be used for theory development (see Figure 1).

We conduct Monte Carlo simulations to demonstrate labGPT's effectiveness in approximating experimental control. In controlled conditions using synthetic data with known ground truth, the results show that labGPT recovers true parameters when textual stimuli included unobserved variations. labGPT results closely approximate the estimates and confidence intervals of an infeasible Oracle benchmark with perfect knowledge of ground truth.

In addition, to illustrate labGPT, we empirically study preference dynamics in wine, where 1,000 consumers each evaluated 32 pairs randomly drawn from nearly 120,000 real tasting notes. Across participant choices for about 50,000 unique wines, the results reveal that preferences for verbally described products are initially malleable. Despite being entirely incidental, wines

presented at the outset become product anchors in shaping later decisions. A follow-up experiment (coffee) replicated the effect. The results support the notion that preference dynamics is affected by the selective accessibility of knowledge about key attributes (e.g., region and varietal) as shaped by product anchors. By systematically sampling stimuli from a large real-world corpus and controlling for non-focal variation across stimuli, labGPT helps facilitate stimulus sampling for products typically described in prose and strengthens the study's internal and external validity. Our proposed labGPT—based design seeks to complement traditional designs for studying consumer behavior.

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. Specifically, hedonic experiences are often described in prose to help consumers anticipate these experiences. 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). labGPT facilitates research endeavors by making such contexts experimentally tractable. For example, our experiments test a novel consumer insight: verbally described products can become anchors in repeated choices. This builds on classic anchoring research, a judgment phenomenon typically studied using simpler stimuli that, it is speculated, may not occur in information-rich environments (Ebbesen and Konecni 1980; Gigerenzer 1991). labGPT can be used to help resolve conflicting theoretical predictions arising from discrepancies in product presentations between simpler, stylized stimuli and detailed, real-world stimuli.

Second, we address the stimulus-sampling problem. To mitigate this problem, scholars recommend that the stimulus sample be enlarged (Westfall, Judd, and Kenny 2015) and treated as

a random factor, viewing each stimulus as one of many possible items drawn from its larger population and sampling stimuli in the same manner as sampling participants (e.g., Baribault et al. 2018; Judd, Westfall, and Kenny 2012). labGPT facilitates these design suggestions for stimulus selection. We presented 1,000 participants with approximately 50,000 real-world wine descriptions randomly sampled from a pool of 119,955 wines produced across 427 wine-growing regions and 708 wine-grape varietals. This sampling approach minimizes the likelihood that the observed effects are driven by extraneous features specific to a particular wine—a concern in conventional experiments that rely on a small set of carefully pretested stimuli designed to elicit specific effects (Pham 2013)—improving the study's internal validity. The observed effects are also likely to generalize to the approximately 70,000 wines that were randomly excluded from the experiment, as the stimuli were chosen randomly and presented in their original form. This directly enhances the external validity of the study findings. Moreover, the existing approach of manual coding (like our analysis in the coffee experiment) would be impractical for our wine experiment. Given that participants each saw randomly sampled wines, manual coding would require approximately 2.5 billion distances. labGPT efficiently computes these distances, accounting for nuisance variables and thereby helping to preserve the research's internal validity. While traditional experiments rely on orthogonal, factorial designs for causal inference, our proposed approach combines large-scale randomization, continuous attribute representation, and statistical nuisance controls to facilitate causal inference. Allowing each participant to be shown a distinct, randomly sampled subset of stimuli leverages the between-item variance that Judd, Westfall, and Kenny (2012) highlight, thereby increasing statistical power without increasing the participant sample size.

Third, our proposed research design can help enhance research objectivity and reproducibility. Traditional designs can involve numerous researcher decisions in (1) designing the experiments, such as stimulus simplification and selection, and (2) data analysis methods. These researcher degrees of freedom can inadvertently compromise research reproducibility. Our framework allows the inclusion of real-world stimuli without abbreviation, facilitates stimulus selection from market coverage, and offers a structured, end-to-end methodological pipeline where key parts of the

analytical plan (e.g., embedding models) can be pre-specified and pre-registered. Not only does it mitigate the stimulus-sampling problem, it overcomes the "garden of forking paths" (Gelman and Loken 2013) problem in some conventional approaches. labGPT supports open science practices like pre-registration, enhancing replicability and the overall credibility of the findings.

Finally, we build on cognitive psychology and neuroscience work showing that LLM embeddings align closely (and linearly) with human brain's own hierarchical codes for speech and meaning (Goldstein et al. 2025). These embeddings can serve as proxy for consumer conceptual knowledge for downstream tasks. However, LLMs are inherently unstructured and high-dimensional, with thousands of unlabelled dimensions. Even though they can predict downstream tasks such as food-health judgments (Gandhi et al. 2022), it remains unclear which specific constructs or semantic features drive these judgments. To facilitate their use for theory testing in consumer behavior, we propose labGPT which combines LLM with a two-stage specialized neural network architecture to develop structured, lower-dimensional, and interpretable representations. Our approach conceptually parallels recent work like Goldstein et al. (2025) to reduce dimensionality but goes further to interpret dimensions for theory testing. labGPT is also designed to adapt to context and different dataset features through automatic tuning (see Web Appendix §F.2).

Limitations, extensions, and directions for future research

Embeddings for images, audios, and videos

labGPT focuses on unimodal (text) product descriptions. We opted to start with text because textual stimuli are prevalent and current AI embeddings are more advanced and accurate with text. However, both our method and existing embedding models can handle multimodal data inputs, such as multimodal product descriptions. These advances suggest promising enhancements to our design's capabilities. As large multimodal models improve, it would be interesting for future research to test consumer behavior theories involving other forms of unstructured stimuli, such as images, audios, videos, and even virtual and augmented reality.

Stimulus sampling method

Large-scale randomization ensures sufficient variation across the attribute space and, on average, approximates balance across attribute levels. In our wine experiment, for each participant and each task, we randomly sampled a wine in each pair from a large real-world corpus. In practice, the same sampling considerations and methods used for participants (e.g., simple random, stratified, cluster sampling) can be applied to stimulus selection. Some sampling methods may be better suited for certain research questions, such as how consumers perceive Old World versus New World wines or the effectiveness of messages that are concise, medium, or lengthy. In these situations with specified bins, researchers can use stratified sampling via wine-provenance or word-count strata for more balanced groupings. We discuss this flexibility in stimulus sampling method and provide guidance in the researcher's guide (Web Appendix §B).

Discovering latent dimensions

A promising direction for future research is to use labGPT to uncover latent dimensions from unstructured texts. To test preference dynamics in our main study, we chose a context in which we know which attributes are likely important to consumers, because this allows us to transform the unstructured tasting notes into a theory-based attribute representation to identify the source of any potential anchoring effects. We can thus more cleanly isolate semantic anchors on theoretically relevant dimensions, in line with our conceptual rationale that products activate select semantic knowledge about the options based on identified key attributes. While we opted for a deductive approach, it is possible to use labGPT to inductively uncover other key latent dimensions. Future research can compare how theory-driven and data-driven dimensions shape preference dynamics.

Consumer contexts with rich, unstructured texts and using multiple dimensions

Relatedly, future research can implement labGPT-based research designs to study other consumer contexts in which unstructured texts are rich and common. For example, one can study what makes online consumer reviews more useful using the procedure in Figure 2 by (a) collecting

a corpus of reviews (step 1), (b) using labGPT to identify the underlying dimensions that affect review usefulness (step 3), and (c) validating the causal impact of these dimensions on review usefulness (steps 2, 4 and 5). Suppose that researchers want to further verify the internal validity of the test, they could develop controlled manipulations of the identified dimensions and use a traditional design in a laboratory study. This parallels our overall empirical strategy for assessing semantic anchoring with labGPT-based wine experiment and the controlled coffee experiment.

Some consumer contexts involve extensive texts describing numerous aspects of the options. For example, insurance and financial products are often presented in rich, extensive texts, and consumers may consider multiple attributes. Such contexts remain challenging to study using conventional methods. Given that labGPT can handle large, diverse product spaces (with high-cardinality categorical variables) through continuous representation, it would be interesting to examine how consumers consider multiple dimensions in decision making using labGPT.

High-level psychological constructs

An important direction for future research is to assess the extent to which labGPT can be applied to study abstract or holistic psychological constructs. In our experiments, we examined more concrete drivers of judgments—specified product attributes—because these attributes are well defined and appropriate for our research question. Because labGPT uses supervised training to disentangle focal constructs from other variations in texts, we speculate that it is better suited when the focal constructs can be reasonably defined and labeled for the supervised training process, even if these constructs are abstract. For instance, it could potentially be used to extract quantifiable aspects of narratives like sentiment arcs (Toubia, Berger, and Eliashberg 2021) or the prevalence of specific thematic elements (Toubia et al. 2019) if these can be reliably labeled in a training set. However, labGPT is likely less suited for capturing purely holistic or highly subjective interpretations where defining clear targets is difficult (e.g., assessing the overall 'persuasiveness' of entirely different narrative structures without breaking them down into measurable components). An area for future research could involve integrating unsupervised topic modeling or clustering techniques

prior to labGPT's supervised stage to help define relevant dimensions within highly abstract texts.

Emerging product categories

Generative AI offers the ability to create descriptions when real-world stimuli are limited, such as in emerging or novel product categories[12], or when study objectives require more controlled descriptions that mimic real products (e.g., the coffee descriptions in our follow-up experiment). Here, researchers could adopt a human-in-the-loop framework, where humans guide the AI in generating stimuli or assist in selecting from AI-generated stimuli (e.g., using platforms like MTurk). This approach could enable researchers to craft realistic stimuli with ease and low cost.

Managerial Implications

The use of real-world stimuli enhances the potential for counterfactual analysis, because inferences drawn from the study relate directly to actual products rather than simplified abstractions. For example, our statistical model maps the preferences of 1,000 consumers in the US, Australia, and New Zealand for all 119,955 wines in our dataset. 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 for different orderings of wine pairings. The estimated preference structure offers potential as a foundation for marketing decision tools that, similar to other structural econometric models, may inform the optimization of marketing strategies. As AI advances, we hope labGPT enables researchers to use naturalistic stimuli in experiments that parallel how these products are presented in the marketplace, helping to advance the study of consumer behavior in information-rich environments.

[12]We illustrate this capability in our researcher's guide (see Web Appendix §B).

REFERENCES

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.

Battleday, Ruairidh M, Joshua C Peterson, and Thomas L Griffiths (2021), "From convolutional neural networks to models of higher-level cognition (and back again)," *Annals of the New York Academy of Sciences*, 1505 (1), 55–78.

Bhatia, Sudeep and Russell Richie (2022), "Transformer networks of human conceptual knowledge." *Psychological Review*.

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

Bisson, Linda F, Andrew L Waterhouse, Susan E Ebeler, M Andrew Walker, and James T Lapsley (2002), "The present and future of the international wine industry," *Nature*, 418 (6898), 696–99.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others (2020), "Language models are few-shot learners," *Advances in neural information processing systems*, 33, 1877–1901.

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

Calder, Bobby J, Lynn W Phillips, and Alice M Tybout (1982), "The concept of external validity," *Journal of consumer research*, 9 (3), 240–44.

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

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

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.

Chang, Hannah H and Michel Tuan Pham (2018), "Affective boundaries of scope insensitivity," *Journal of Consumer Research*, 45 (2), 403–28.

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.

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.

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.

Griffiths, Thomas L, Mark Steyvers, and Joshua B Tenenbaum (2007), "Topics in semantic representation." *Psychological review*, 114 (2), 211.

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.

Landauer, Thomas K and Susan T Dumais (1997), "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, 104 (2), 211.

Landauer, Thomas K, Peter W Foltz, and Darrell Laham (1998), "An introduction to latent semantic analysis," *Discourse processes*, 25 (2-3), 259–84.

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.

Laverghetta Jr, Antonio, Animesh Nighojkar, Jamshidbek Mirzakhalov, and John Licato (2022), "Predicting human psychometric properties using computational language models," in *Quantitative psychology: The 86th annual meeting of the psychometric society, virtual, 2021*, Springer,

151–69.
Levesque, Hector J (1986), "Knowledge representation and reasoning," *Annual review of computer science*, 1 (1), 255–87.
Lutz, Richard (2018), *2018 ACR fellow address: On relevance*, (T. W. Andrew Gershoff Robert Kozinets, ed.), Duluth, MN: Association for Consumer Research, 14–22.
Lynch Jr, John G (1982), "On the external validity of experiments in consumer research," *Journal of consumer Research*, 9 (3), 225–39.
Lynch Jr, John G, Joseph W Alba, Aradhna Krishna, Vicki G Morwitz, and Zeynep Gürhan-Canli (2012), "Knowledge creation in consumer research: Multiple routes, multiple criteria," *Journal of Consumer Psychology*, 22 (4), 473–85.
Lynch Jr, John G and Dan Ariely (2000), "Wine online: Search costs affect competition on price, quality, and distribution," *Marketing science*, 19 (1), 83–103.
MacNeil, Karen (2015), *The wine bible*, Workman Publishing.
Markman, Arthur B (2013), *Knowledge representation*, Psychology Press.
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.
Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013), "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, 26.
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.
Mukhopadhyay, Anirban, Priya Raghubir, and S Christian Wheeler (2018), "Judgments of taste and judgments of quality."
Mussweiler, Thomas (2003), "Comparison processes in social judgment: Mechanisms and consequences." *Psychological review*, 110 (3), 472.
Mussweiler, Thomas and Fritz Strack (2001), "The semantics of anchoring," *Organizational behavior and human decision processes*, 86 (2), 234–55.
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.
Peterson, Joshua C, Joshua T Abbott, and Thomas L Griffiths (2018), "Evaluating (and improving) the correspondence between deep neural networks and human representations," *Cognitive science*, 42 (8), 2648–69.
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.
Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and others (2019), "Language models are unsupervised multitask learners," *OpenAI blog*, 1 (8), 9.
Robinson, Jancis (2015), *The oxford companion to wine*, American Chemical Society.
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.
Sharot, Tali, Cristina M Velasquez, and Raymond J Dolan (2010), "Do decisions shape preference? Evidence from blind choice," *Psychological science*, 21 (9), 1231–35.
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.

Storrs, Katherine R, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte (2021), "Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting," *Journal of Cognitive Neuroscience*, 33 (10), 2044–64.

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.

Toubia, Olivier, Garud Iyengar, Renée Bunnell, and Alain Lemaire (2019), "Extracting features of entertainment products," *Journal of Marketing Research*, 56 (1), 18–36.

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.

Yoon, Song-Oh and Itamar Simonson (2008), "Choice set configuration as a determinant of preference attribution and strength," *Journal of Consumer Research*, 35 (2), 324–36.