Check for updates

The RAND Journal of Economics

ORIGINAL ARTICLE OPEN ACCESS

Demand Estimation with Text and Image Data

Giovanni Compiani[1] | Ilya Morozov[2] | Stephan Seiler[3]

[1]Booth School of Business, University of Chicago, Chicago, Illinois, USA | [2]Kellogg School of Management, Northwestern University, Evanston, Illinois, USA | [3]Imperial College Business School, Imperial College London & CEPR, London, UK

Correspondence: Giovanni Compiani (gio1compiani@gmail.com)

Received: 30 September 2025 | **Revised:** 14 February 2026 | **Accepted:** 3 March 2026

Keywords: demand estimation | unstructured data | deep learning

ABSTRACT

We propose a demand estimation approach that leverages unstructured data to infer substitution patterns. Using pre-trained deep learning models, we extract embeddings from product images and textual descriptions and incorporate them into a mixed logit demand model. This approach enables demand estimation even when researchers lack data on product attributes or when consumers value hard-to-quantify attributes such as visual design. Using a choice experiment, we show this approach substantially outperforms standard attribute-based models at counterfactual predictions of second choices. We also apply it to 40 product categories offered on Amazon.com and consistently find that unstructured data are informative about substitution patterns.

1 | Introduction

Many problems in economics and marketing—such as merger analysis, tariff evaluation, and optimal pricing—require estimating demand for differentiated products. A standard approach has been to estimate demand models that capture substitution through the similarity of product attributes. While common, this approach faces two practical challenges. First, researchers rarely observe all choice-relevant attributes. Instead, they rely on third-party data in which attributes are chosen based on unknown criteria, or gather their own data, subjectively choosing which attributes to collect. Second, consumers often consider visual design and functional benefits of products, dimensions that are difficult to capture through observed attributes.[1]

In this paper, we show how researchers can incorporate unstructured data—such as product images, descriptions, or reviews—in demand estimation to recover substitution patterns. Using pre-trained deep learning models, we transform images and texts into vector representations (*embeddings*) and apply Principal Component Analysis (PCA) to reduce dimensionality and capture the main dimensions of product differentiation. We incorporate these principal components into a logit model with random coefficients, treating them as researchers usually treat observed product attributes. Lastly, we implement a model selection algorithm to choose the type of unstructured data and deep learning model that extract the strongest signal of substitution.

This approach lets researchers incorporate hard-to-quantify product attributes, such as visual design in images and functional benefits from text, while avoiding subjective choices about which attributes to collect. It relies on product images and textual descriptions, which are widely available even in markets where collecting product attributes is challenging. Given the prevalence of unstructured data, the approach offers a valuable addition to the toolbox of empirical researchers. To ease adoption, we provide a Python package, *DeepLogit*, alongside this paper.[2]

We validate this approach empirically. A key challenge is that substitution patterns are unobserved, so there is no ground truth to test model predictions. An ideal dataset would let us estimate demand precisely from consumers' choices and then evaluate how well the model predicts an empirical measure of substitution not used in estimation. To achieve this, we design an experiment in which participants choose a book from a list

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *The RAND Journal of Economics* published by Wiley Periodicals LLC on behalf of The RAND Corporation.

of options, and we elicit both first and second choices. We randomize prices and rankings of all options. This variation lets us estimate substitution patterns using only first choices and without instruments, while reserving second choice data for validation. We then test how well the model predicts second choices counterfactually.[3] Because second choices reveal which product a consumer substitutes to if their preferred option is unavailable, predicting them directly validates the model' s ability to recover substitution patterns.

Using these experimental data, we show that both text and images contain useful information about substitution. Intuitively, book covers visually convey information about genre, while descriptions and reviews capture plot details that identify similar books even within the same genre. Our model selection algorithm chooses a review-based mixed logit. This selected model improves counterfactual predictions of second choices relative to a standard attribute-based mixed logit. In fact, the improvement relative to that model is on par with the improvement that the attribute-based mixed logit achieves over plain logit without random coefficients. Thus, using unstructured data alone, our approach recovers substitution patterns substantially better than standard attribute-based models.

We further validate our approach using observational data from 40 product categories offered on Amazon.com, including groceries, pet food, office supplies, beauty products, electronics, video games, and clothing. We combine online purchases from the Comscore Web Behavior panel with product images, titles, descriptions, and reviews collected from Amazon's product pages. We show that text and images contain useful signals about substitution in all studied categories, highlighting that our approach is broadly applicable across markets. We find that text performs best in some categories where images might seem ex ante more relevant (e.g., clothing), and vice versa. Thus, it is helpful to collect both text and image data even when only one seems relevant a priori.

Our paper contributes to the extensive literature on demand estimation. We build on papers that model substitution through the heterogeneity of preferences over observed product attributes (Berry et al. 1995; McFadden and Train 2000; Nevo 2001; Berry et al. 2004). We show how researchers can incorporate unstructured data into these standard models to improve estimation of substitution patterns. Our key contribution is to validate this approach extensively using both experimental and observational data.

We also build on a vast computer science literature that transforms unstructured data into lower-dimensional representations (Krizhevsky et al. 2012; Mikolov et al. 2013; Goodfellow et al. 2016) and a growing economics and marketing literature that applies pre-trained deep learning models (Battaglia et al. 2024; Dew 2024). Unlike this prior work, which focuses on predicting observed outcomes, we aim to predict substitution patterns counterfactually—a crucial task for applications like merger simulations and optimal pricing. Because substitution patterns are usually unobserved, we cannot apply standard cross-validation and instead need to design an experiment to measure them. To our knowledge, we are the first to use such an experiment to validate demand models.[4]

Several other papers leverage text and image data in demand estimation. Quan and Williams (2019) and Lee (2024) incorporate product images and textual descriptions as mean utility shifters. Our approach models utility covariances and can thus capture substitution patterns more flexibly. Han and Lee (2025) use image data to estimate substitution among fonts and study the effects of copyright protection policies. We contribute by validating our approach on ground truth substitution patterns, comparing counterfactual prediction performance across several embedding models and types of unstructured data, and providing practical guidance for demand estimation. Sisodia et al. (2024) extract interpretable product attributes from images and use them to estimate preferences in a choice-based conjoint experiment. Because our approach is easier to implement and scales better across categories, it is preferable when researchers do not need the extracted attributes to be interpretable.

An alternative way to learn substitution patterns is to use survey data. For example, Dotson et al. (2019) ask survey participants to rate each product image and then incorporate rating correlations as shifters of utility correlations into a demand model. Similarly, Magnolfi et al. (2022) elicit relative rankings from survey participants (e.g., "Product A is closer to B than to C"), generate embeddings from these rankings, and use them in a mixed logit model. Our approach complements these survey-based methods but has the advantage of using only widely available text and image data, thereby avoiding the need for costly category-specific surveys.

Our approach is well-suited for empirical applications that require accurate and flexible estimates of product substitution. This includes evaluating how horizontal mergers (Nevo 2000; Federal Trade Commission 2022), new product launches (Hausman 1994; Petrin 2002), corrective taxes (Allcott et al. 2019; Seiler et al. 2021), and trade restrictions (Goldberg 1995; Berry et al. 1999) influence consumers' choices and welfare through changes in assortments and prices. It can also be used to study how multi-product firms should optimize prices and promotions (Hoch et al. 1995; Vilcassim and Chintagunta 1995). The approach avoids the need to collect category-specific attributes, making it easier to estimate demand across many categories (Döpper et al. 2024).[5] Finally, this approach might be particularly useful for estimating demand for creative goods—such as books, art, and movies—where unstructured data is likely to capture product differentiation more fully than observed attributes alone.

2 | Proposed Approach

The proposed approach involves three steps: (1) extracting embeddings from images and texts, (2) reducing the dimensionality of these embeddings using PCA, and (3) including the resulting principal components into a standard attribute-based logit model with random coefficients.

2.1 | Step 1: Extracting Embeddings From Texts and Images

To extract information from product images, we compute low-dimensional representations, *embeddings*, using pre-trained deep

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

learning models. We use four convolutional neural networks: VGG19 (Simonyan and Zisserman 2015), ResNet50 (He et al. 2016), InceptionV3 (Szegedy et al. 2016), and Xception (Chollet 2017).[6] The key advantage of using pre-trained models is that it reduces the computational burden of estimation while leveraging models that were trained on large-scale datasets and are known for their strong performance in image classification tasks (Russakovsky et al. 2015). Because these models perform well at distinguishing visually similar objects, we expect embeddings to capture the key visual features that differentiate products. Rather than committing to a single model, we perform model selection to determine which one best captures substitution patterns, as described below.

We also extract information from texts, including product titles, descriptions, and reviews. We use a bag-of-words model, which represents text as fixed-length vectors based on word counts, as well as a variation with a TF-IDF vectorizer.[7] While these count models are relatively simple, we view them as useful benchmarks because they can still detect attributes mentioned in textual product descriptions. In addition, we employ two pre-trained deep learning models: Universal Sentence Encoder (USE) (Cer et al. 2018), and BERT Sentence Transformer (ST) (Reimers and Gurevych 2019).[8] Both USE and ST produce semantically meaningful sentence embeddings and achieve excellent performance on semantic textual similarity benchmarks (Cer et al. 2017). As a result, they can identify when sellers or consumers describe the same product attributes and functional benefits using similar language, even if the exact words differ.

2.2 | Step 2: Generating Principal Components

Although embeddings compress texts and images into lower-dimensional representations, they remain high-dimensional compared to the number of attributes typically included in demand models. For example, incorporating 512-dimensional VGG19 embeddings into a random coefficients logit model would require prohibitively costly numerical integration over all dimensions. We apply PCA to further reduce dimensionality (Backus et al. 2021).

PCA is appealing for several reasons. Because embeddings are trained on general-purpose datasets to classify images into broad categories, they contain variation that distinguishes each category from one another (e.g., laptops from tablets). By applying PCA within each category, we downweight this common variation and focus on differences across products within that category. Additionally, the resulting principal components are orthogonal to each other, which avoids multicollinearity issues and thus simplifies estimation of logit demand models with random coefficients.

2.3 | Step 3: Including Principal Components into a Demand Model

We include principal components into a standard mixed logit model (Berry et al. 1995;2004), estimating a separate random coefficient for each included component. Since each deep learning model and data type produces different principal components, a key question is which to include in the demand model.[9] We perform model selection using in-sample *AIC*, which balances model fit and complexity and can be interpreted as approximating expected out-of-sample predictive performance (Akaike 1998). In our experimental results, in-sample *AIC* strongly correlates with counterfactual performance, justifying its use for model selection (see Section 3.3 for a discussion and comparison to *BIC* and cross-validation).

Algorithm 1 describes our model selection procedure. We define the *candidate set* as the variables for which random coefficients may be included. In our applications, this set includes price and the first P principal components. We first estimate the plain logit model without random coefficients ($K = 0$) and record its *AIC*. For each $K \geq 1$, we estimate all models with random coefficients on K candidate variables and retain the model with the lowest *AIC*. If this *AIC* is lower than that of the previously selected model, we update the selected specification and increase K; otherwise, we stop and select the previous model. We increase K until *AIC* no longer improves. We repeat this procedure for each specification—defined by data type and embedding model (e.g., *USE Reviews*)—and choose the specification with the lowest overall *AIC*.

This algorithm avoids costly combinatorial search over all possible specifications. At the same time, it ensures that we do not rely too heavily on the default ordering of principal components, which reflects the explained variance of embeddings but need not match how predictive they are of preferences. For example, substitution patterns may be driven by the second and fourth principal components. In this case, our approach may select a model with random coefficients on those components, but not on the first or third.

Algorithm 1: Model Selection

Input:
- Products $j = 1,\ldots,J$
- Specifications $m = 1,\ldots,M$ (all combinations of unstructured data and pre-trained text or image models)
- Maximum number of principal components P (researcher's choice)
- Demand model with utility: $u_{ij} = \delta_j + \alpha_i \text{price}_{ij} + \theta_i' PC_j + \varepsilon_{ij}$ (δ_j = product fixed effects)

For each specification m **do**
  Extract embeddings $e^{(m)} = (e_1^{(m)},\ldots,e_J^{(m)})$
  Extract principal components $PC_1^{(m)},\ldots,PC_P^{(m)}$ from embeddings $e^{(m)}$ using PCA
  Initialize model selection at $K \leftarrow 0$ (zero random coefficients)
  Set BestSpec$^{(m)}$ as *Plain Logit* and BestAIC$^{(m)}$ as the corresponding *AIC*
  while BestAIC$^{(m)}$ decreases **do**
    $K \leftarrow K + 1$
    Estimate mixed logit models with all possible combinations of random coefficients on subsets of K variables from the candidate set $\mathcal{R} = \{\text{price}, PC_1^{(m)},\ldots,PC_P^{(m)}\}$
    Find subset $R_K^* \subseteq \mathcal{R}$ that minimizes *AIC* at AIC_K^*
    if $AIC_K^* <$ BestAIC$^{(m)}$ **do**
      BestSpec$^{(m)} \leftarrow R_K^*$: update the best specification
      BestAIC$^{(m)} \leftarrow AIC_K^*$: update the lowest *AIC*
    end if
  end while
end for
Choose the best-fitting specification $m^* = \arg\min_m \text{BestAIC}^{(m)}$

2.4 | Why Does This Approach Work?

The key feature of our approach is that it uses text and images as proxies for product attributes that shape substitution. Take tablets, for example: product titles may reveal brand, screen size, and camera resolution (e.g., "Apple iPad 10.2-inch 12MP camera"); seller descriptions may highlight whether the tablet is suitable for drawing or gaming; and consumer reviews may

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

mention that a tablet is durable and child-friendly. Similarly, photos may showcase design features like color and casing style. Our approach extracts this information in an automated way and uses it to estimate substitution patterns.

We interpret the relationship between embeddings and substitution as correlational. In particular, our approach does not assume that consumers directly consider product images or textual descriptions when making choices. Under this interpretation, our approach is well-suited for counterfactuals where embeddings can be held fixed, such as optimal pricing or merger simulations. By contrast, it is less appropriate for counterfactuals that alter product design or positioning and, in turn, the embeddings themselves. In such cases, researchers are required to estimate the causal effect of embeddings on demand. Section 5.2 provides additional discussion.

2.5 | Advantages Over Standard Methods

Our approach has several advantages relative to the standard attribute-based methods. First, researchers typically select a limited set of attributes based on their prior knowledge of the market, often those that are easiest to quantify, or rely on attributes supplied by data providers. Our approach avoids these subjective choices by automatically extracting information about product substitution from unstructured data. Second, this approach captures visual design and functional benefits, which may drive substitution but are difficult to capture through observed attributes. Lastly, we circumvent the need to collect category-specific attributes, which makes our approach more scalable. This is valuable for researchers seeking to estimate demand and study competition and pricing across many product categories.

When some data on observed attributes are available, researchers can incorporate them into our approach. For example, they can expand the set of candidate variables in Algorithm 1 to include the observed attributes themselves or, when the number of attributes is large, their principal components. In our experiment, adding observed attributes does not improve counterfactual predictions, indicating that the information in the attributes is subsumed by the unstructured data. See Sections 3.3 and 5.1 for details.

3 | Validation With Experimental Data

Next, we validate our approach empirically. A key challenge is that substitution patterns are unobserved, so there is typically no ground truth against which to test model predictions. To address this challenge, we design a choice experiment that provides such a ground-truth measure of substitution.

3.1 | Experiment Design

3.1.1 | Choice Tasks

Each participant completes two choice tasks. In the first choice task, a participant chooses one book from a set of ten alternatives. We instruct participants to choose books they would purchase if faced with this selection in a real bookstore. We limit the number of options to ten to ensure participants pay attention to most options, which allows us to abstract away from limited consideration. After a participant selects a book, we remove it from the list. The participant then proceeds to a second choice task, where they choose among the remaining nine books.

Figure 1 shows the choice task as presented to participants. To give all models a fair chance at capturing substitution, we display attributes (author, year, genre, pages), cover images, and texts (plot description and five reviews), collected from Amazon product pages.[10] We take all books from Amazon's bestseller lists, sampling them in a way that does not bias our model comparisons in favor of any particular specification (see Appendix B for details). Lastly, we randomize book rankings and prices across participants, keeping them fixed across choice tasks for the same participant.[11] This variation allows us to cleanly estimate substitution patterns from only first choices and reserve second choice data for validation (Berry and Haile 2024).

The book category is well-suited for our analysis because participants are likely to consider both structured attributes (e.g., genre) and unstructured information (e.g., plot descriptions). It is also unclear a priori whether substitution patterns are better predicted by images or texts—while texts describe rich plot details, cover images contain clear cues of genres, and whether the book is fiction or non-fiction (see Section 3.3 for further discussion).

3.1.2 | Recruitment and Sample Selection

We recruited 10,775 participants from the online platform Prolific between June 14 and June 27, 2024. This sample size is close to our pre-registered target of 10,000, determined from power calculations in a pilot experiment.[12] Following the pre-registered sample selection criteria, we excluded 14% of all participants because they failed comprehension questions, did not complete the survey, or spent less than one minute on the study, leaving a final sample of 9265 participants.

Because the choice tasks were hypothetical and not incentivized, it is important to verify that participants made meaningful selections. In Appendix C, we show that participants did not rush through the survey, responded to changes in book rankings and prices, and made choices consistent with their self-reported genre preferences.

3.2 | Model Specifications and Estimation Results

We compare our approach against two benchmark models: the plain logit model and a mixed logit model with attributes. All three models fall into a framework where the indirect utility of participant i from book j is

$$u_{ij} = \beta_i' x_j + \theta_i' PC_j + \gamma \cdot \text{rank}_{ij} + \alpha_i \cdot \text{price}_{ij} + \delta_j + \varepsilon_{ij}. \quad (1)$$

Here x_j are observed book attributes: genre dummies, publication year, and length in pages—all attributes participants observe in the choice tasks; PC_j are principal components extracted from embeddings, price_{ij} and rank_{ij} are the price and position of book j in participant i's choice task, δ_j are product fixed effects,

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

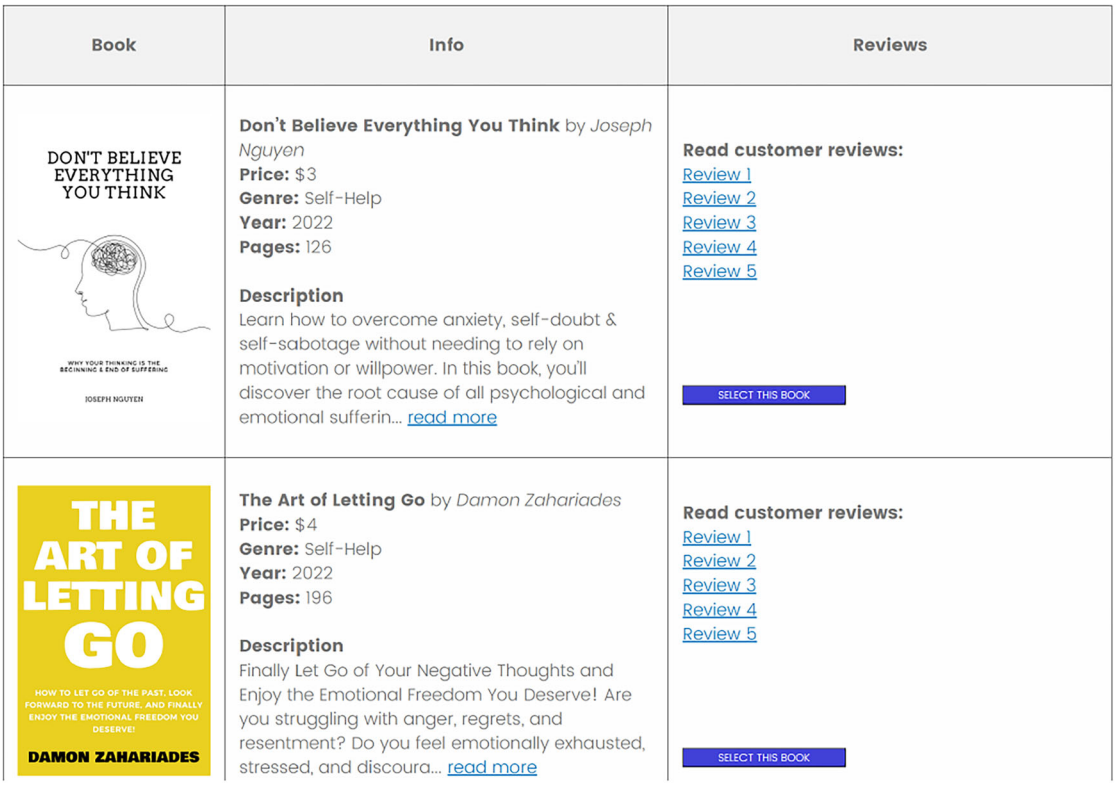


FIGURE 1 | Example of a choice task in our experiment.
Note: The screenshot displays the top portion of the page as it appeared to participants.

and ε_{ij} are i.i.d. taste shocks following a Type I Extreme Value distribution. We include rank_{ij} in the model because our experiment induces random variation in the placement of books.[13] This variation gives us an additional exogenous shifter that helps us identify substitution patterns, analogous to price variation.

We estimate three demand models:

1. **Mixed Logit With Principal Components**. We include principal components PC_j, omit observed attributes x_j, and assume $\alpha_i \sim N(\bar{\alpha}, \sigma_\alpha)$ and $\theta_i \sim N(0, \Sigma_\theta)$ with diagonal Σ_θ.[14] We perform model selection as detailed in Algorithm 1 in Section 2.3, letting the set of candidate variables include price and the first $P = 6$ principal components.[15]
2. **Mixed Logit With Attributes**. We include observed attributes x_j, omit principal components PC_j, and assume $\alpha_i \sim N(\bar{\alpha}, \sigma_\alpha)$ and $\beta_i \sim N(0, \Sigma_\beta)$ with diagonal Σ_β. We choose the lowest *AIC* specification among all possible combinations of random coefficients on price, publication year, length in pages, and genre.
3. **Plain Logit**. We include neither attributes x_j nor principal components PC_j, estimating only a constant price coefficient $\alpha_i = \alpha$, rank coefficient γ, and fixed effects δ_j.

Table 1 compares in-sample *AIC*. The best mixed logit specification with attributes includes random coefficients on the number of pages and year, reducing *AIC* by 16.0 relative to plain logit. The best model with unstructured data, *Reviews USE*, puts random coefficients on the first two principal components. This model fits better than any other considered specification, reducing *AIC* by 24.8 relative to plain logit and by 8.8 relative to the best attribute-based logit. While this superior fit is reassuring, it does not guarantee better counterfactual performance. We therefore turn to predicting second choices counterfactually.

3.3 | Validation Using Second Choice Data

We assess how well different demand models predict second choices counterfactually. Because second choices reveal which books participants choose when their preferred option is unavailable, predicting them accurately requires a model that captures substitution patterns well. Further, predicting second choices is directly relevant in antitrust where second choice *diversion ratios* are often used as measures of substitutability and price competition (Conlon and Mortimer 2021).

We first estimate each model via Maximum Likelihood Estimation (MLE) using only first choices. We then use the estimated model to predict counterfactual second choice diversion ratios, $\hat{s}_{j\to k}$, defined as the probability that the participant chooses product k in the second choice task conditional on having chosen book j in the first choice task. We compare these predictions with diversions observed in the data, $s_{j\to k}$, computing *RMSE* as:

$$RMSE = \sqrt{\frac{1}{J(J-1)} \sum_j \sum_{k \neq j} \left(s_{j\to k} - \hat{s}_{j\to k}\right)^2} \tag{2}$$

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 | Comparison of models in terms of in-sample *AIC* on first choices.

Model	Random Coefficients	*AIC*	Δ*AIC*
Plain logit	None	41006.7	0.0
Mixed logit with attributes	Pages and year	40990.7	−16.0
Mixed logit with images (InceptionV3)	Price, PC1, and PC6	40990.5	−16.2
Mixed logit with titles (ST)	PC1 and PC5	40992.3	−14.4
Mixed logit with descriptions (USE)	PC1 and PC5	40986.4	−20.3
Mixed logit with reviews (USE)	PC1 and PC2	40981.9	−24.8

Note: The second column shows variables that have random coefficients in the selected specification. The last column shows the *AIC* reduction relative to plain logit.

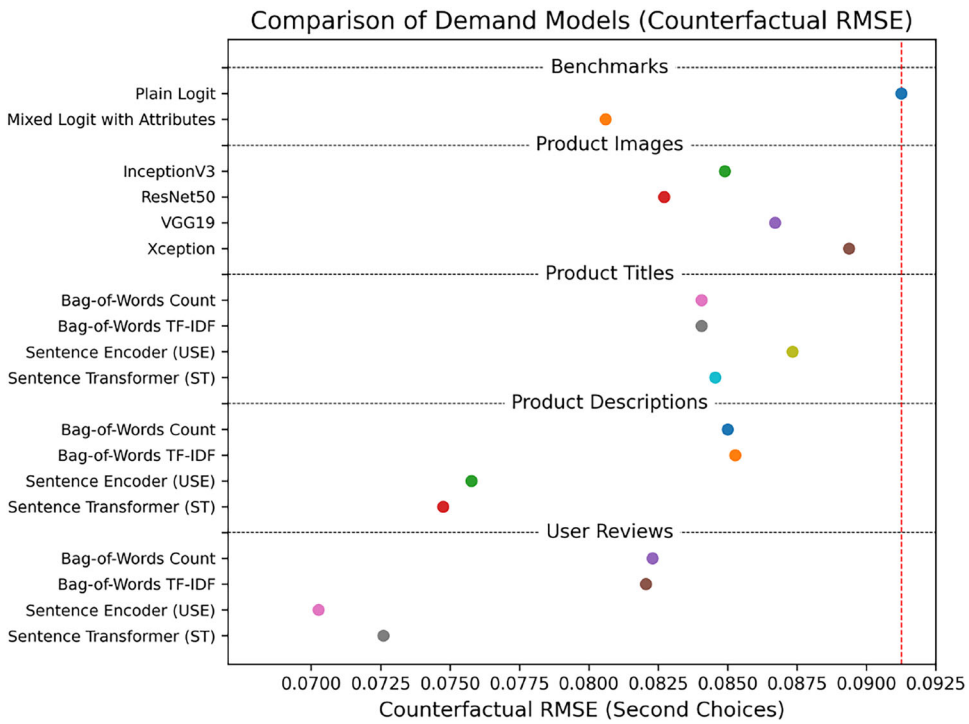


FIGURE 2 | Comparison of models in terms of counterfactual *RMSE* on second choices.
Note: The two benchmarks in the first panel are the plain logit without random coefficients and the lowest-*AIC* mixed logit with random coefficients on observed attributes. The remaining specifications correspond to mixed logit models with random coefficients on principal components extracted from image or text embeddings.

Since we do not use second choice data in estimation, lower *RMSE* indicates better performance at counterfactual predictions. Appendix D provides further details on how we compute diversions $s_{j\to k}$ and $\hat{s}_{j\to k}$.

Figure 2 summarizes the validation results, while Table A1 reports *AIC* and *RMSE* values for all estimated specifications. The top panel in Figure 2 shows two benchmarks: the plain logit and the lowest-*AIC* mixed logit with attributes. The mixed logit with attributes reduces *RMSE* by 11.7%, while our best-fitting *Review USE* model reduces *RMSE* by 23% relative to plain logit, significantly outperforming the mixed logit with attributes. This result illustrates the value of our approach. We chose books for this study expecting observed attributes like genre to predict substitution well. Yet, by using unstructured data, we can match the counterfactual performance of the mixed logit with attributes and even further reduce *RMSE* by 14%. This shows that our approach recovers substitution patterns better than standard attribute-based methods in this dataset.

Although the model with product reviews performs best, in general, performance varies widely across specifications. Below, we discuss why some specifications recover substitution patterns better than others.

3.3.1 | Images

All four image models outperform the plain logit model, with the selected (i.e., lowest-*AIC*) model, InceptionV3, reducing *RMSE* by 7.0% relative to plain logit. To understand why images predict

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

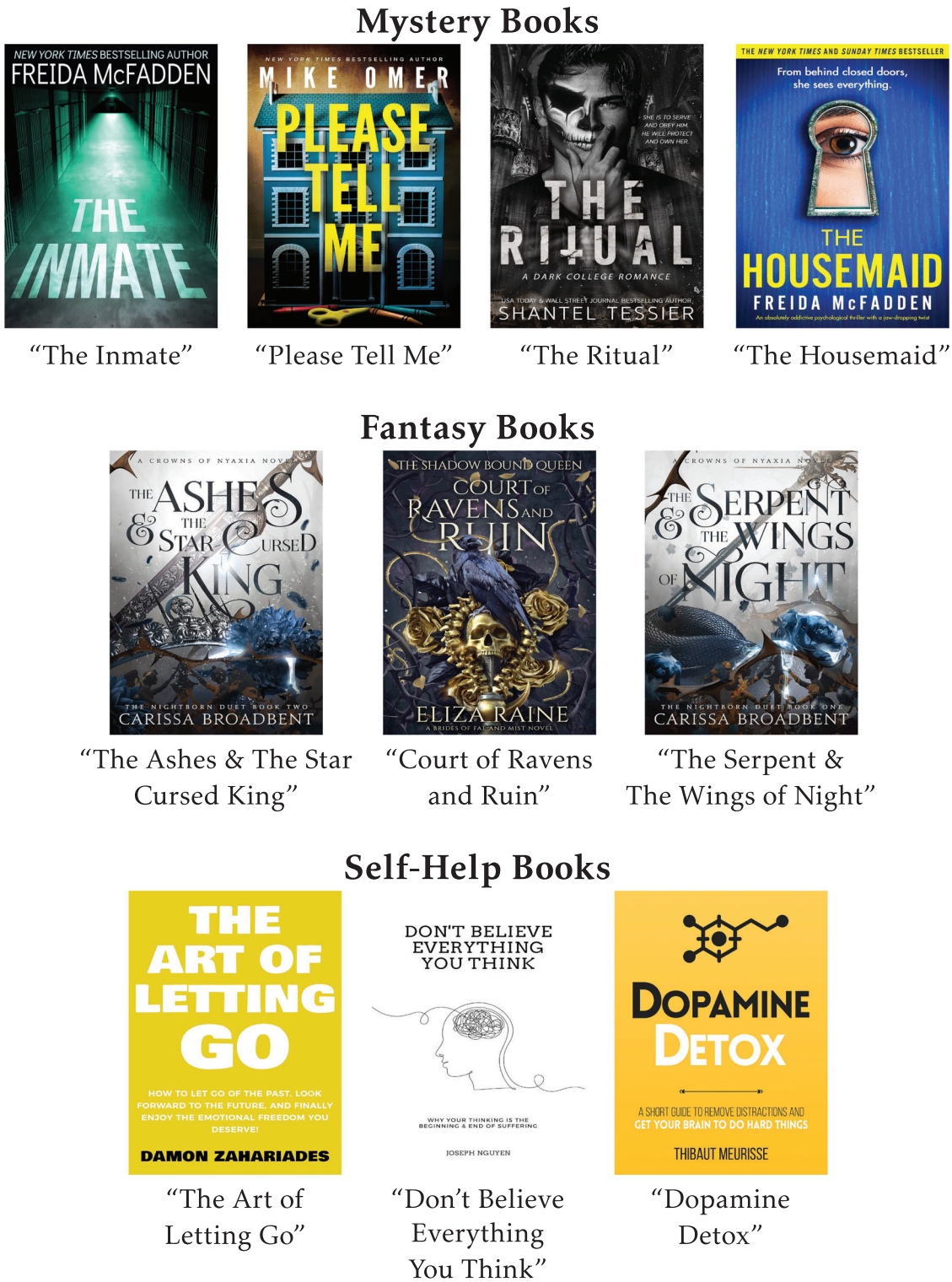


FIGURE 3 | Ten books used in our experiment.

substitution, consider the book covers used in our study (see Figure 3). Within the same genre, book covers often share similar design elements. For example, the covers of all three fantasy books use dark, muted color palettes with metallic accents, include symbolic elements such as skulls and swords that convey danger or peril, and feature natural objects like twisted vines and golden roses. Similarly, the self-help books have minimalistic layouts and consistent color schemes, such as black-and-white text on yellow backgrounds. Thus, image embeddings partly encode books' genres, which correlate with substitution patterns.

3.3.2 | Texts

All text models outperform the plain logit, and their relative performance highlights several notable patterns.

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

First, performance improves as we move from simple bag-of-words models to the more advanced USE and ST models. This is particularly true for descriptions and reviews, which contain detailed information about book plots but do not always use the same words or phrases. Therefore, extracting substitution patterns from text requires natural language models that can accurately measure semantic similarity.

Second, performance improves as we include richer text data. For instance, relative to the plain logit, the USE model reduces *RMSE* by 4% using titles, 17% using descriptions, and 23% using reviews. This makes intuitive sense. Take mystery books, for example. Titles may contain subtle genre cues, hinting at characters in confined situations (*Housemaid*, *Inmate*). Descriptions reveal further plot details, such as hidden identities, past relationships, and betrayals. Lastly, customer reviews contain the richest information about a book's style and pacing: for instance, readers may praise cliffhangers and intriguing twists, or critique pacing issues like slow starts or rushed endings.

3.3.3 | Extension: Combining Data Types

If text, images, and attributes provide distinct signals of substitution, combining them might improve fit and counterfactual performance. To explore this, we start from the selected *Review USE* model and attempt to extend it in two ways: (1) by adding a combination of observed product attributes with random coefficients, or (2) by adding a combination of principal components from InceptionV3, the lowest-*AIC* image model. In each case, we evaluate all subsets of added variables and estimate random coefficients for these added variables, in addition to those already included in the selected *Review USE* model.[16] We find that in all evaluated specifications, the variances of the added random coefficients are zero. As a result, *AIC* increases and *RMSE* does not improve. This result suggests information is highly correlated across data types, with text providing the strongest signal of substitution.

3.3.4 | What do Principal Components Capture?

To better understand the variation captured by embeddings, in Figure 4 we show book locations in the space of the two principal components selected into our best-performing *Review USE* model. These principal components align with intuitive substitution patterns: the first one (horizontal axis) separates non-fiction on the left from fiction on the right, while the second (vertical axis) further distinguishes science fiction from mystery. Crucially, the variation in these principal components goes beyond separating genres. For example, *The Housemaid* and *The Inmate* are by the same author, and *The Serpent & The Wings* and *The Ashes & The Star-Cursed King* are from the same book series. The two principal components detect this similarity from consumers' reviews of these books. As evident from *RMSE* comparisons in Figure 2, this additional variation helps recover substitution patterns better and improves counterfactual predictions.

3.3.5 | Implied Substitution Patterns

To illustrate how our approach improves predictions of substitution patterns, in Table 2 we compare predicted second choice probabilities with their counterparts observed in the experimental data for three of the books.

Panel A examines substitution patterns for the self-help book *Dopamine Detox*. The two other self-help books are, by far, its closest substitutes in the data. The plain logit model misidentifies these substitutes, incorrectly predicting that people would switch to books *Please Tell Me* and *The Inmate*—two popular books with the largest market shares. The attribute-based mixed logit correctly identifies *Don' t Believe Everything You Think* as the closest substitute but mispredicts the second one, likely due to its over-reliance on estimated fixed effects. By contrast, our review-based model is the only one that correctly predicts all five closest substitutes in the correct order.

Panel B shows a similar example with the substitutes for the mystery book *Please Tell Me*. The plain logit model mispredicts second choice probabilities, incorrectly suggesting that the second-closest substitute is a self-help book. By contrast, both the attribute-based mixed logit and our review-based logit capture strong within-genre substitution, correctly identifying the top three closest substitutes. Additionally, the review-based model recognizes that *The Inmate* and *The Housemaid* have significantly higher second choice probabilities than the third-closest substitute—an insight that the attribute-based model misses.

These examples illustrate that, beyond reducing *RMSE* and predicting second choice probabilities more accurately on average, our approach can learn *which* products are the closest substitutes.

Despite these favorable examples, our approach does not always accurately capture substitution. Panel C shows an example where the review-based model misidentifies the closest substitutes for the fantasy book *The Ashes & The Star-Cursed King*. In fact, all three models fail, incorrectly predicting mystery and self-help books as the closest substitutes. These deviations from observed second choices suggest that there is not enough variation in the first-choice data to reliably estimate substitution patterns for some alternatives.

Finally, to illustrate how estimated substitution patterns affect counterfactuals, we perform a stylized merger simulation in Appendix E. We find that the selected review-based model predicts substantially different post-merger prices than benchmark models.

3.3.6 | Relation Between In-Sample and Counterfactual Performance

Although we select models based on first-choice *AIC*, our experimental data allows us to verify whether this selection algorithm indeed chooses specifications with the best counterfactual performance. Two findings support our choice of *AIC* as a model selection criterion. First, across all specifications with principal components considered by our model selection algorithm, the correlation between first-choice *AIC* and counterfactual second choice *RMSE* is 0.78. Second, *AIC* selects the

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

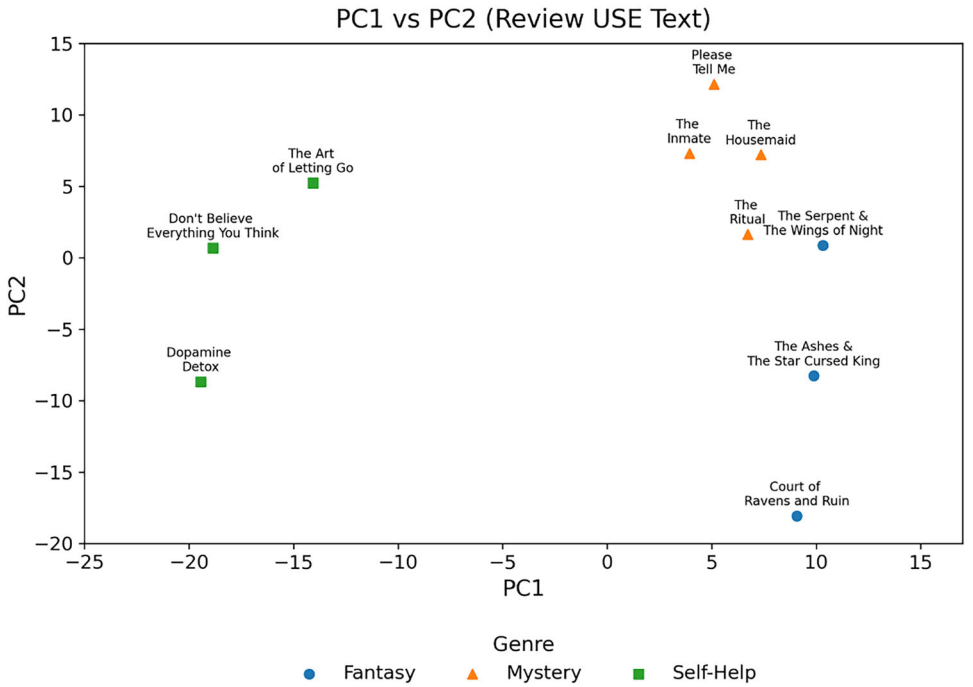


FIGURE 4 | Book locations in the space of selected principal components (Review USE model).

specification that has the lowest second choice *RMSE* across all considered specifications.

Lastly, using *BIC* instead of *AIC* selects the same specification, as does a five-fold cross-validation procedure on the first choices. This reassures us that the key results are robust to the choice of model selection method.

4 | Application to Online Retail Data

Next, we apply our approach to choice data from several online markets. Our first goal is to show that this approach can be applied broadly across categories without being tailored. Our second goal is to determine which types of unstructured data best predict substitution patterns in various categories and to offer practical guidance on what data researchers should collect for demand estimation.

4.1 | Data

We use purchase data from the 2019-2020 Comscore Web Behavior Panel. Specifically, we use the dataset constructed by Greminger et al. (2023), who classify over 12 million unique products from Amazon.com—browsed or purchased by Comscore panelists—into narrowly defined categories. This dataset also matches purchases with daily product price histories obtained from the third-party database Keepa.com.[17] We focus on Amazon.com due to the large volume of Amazon purchases in the Comscore dataset.

We combine purchase data with image and text data we collected from Amazon product detail pages (Figure 5). We use the default product photo for image embeddings. For text embeddings, we use product titles, product descriptions (i.e., bullet points describing the product), and 100 most recent reviews for each product. In estimation, we treat text and image embeddings as fixed over time. While sellers could theoretically modify images and textual descriptions to reposition their products, we find that product images, titles, and descriptions very rarely change over time (see Appendix G).

We apply our method to 40 product categories spanning clothing (shirts, blouses, underwear, sleepwear), household goods (paper towels, trash bags, batteries), office supplies (pens, markers, printing paper), groceries (tea, coffee, bottled water), pet food (wet food, treats), electronics (tablets, monitors, headphones, memory cards, media players), and video games (PC, Nintendo, Xbox, PlayStation). We select the 15 most-purchased products in each category and retain only those categories where these products collectively account for at least 2000 purchases. These selection criteria ensure that we can estimate product fixed effects precisely, which helps capture time-invariant differences across products, and that we have sufficient variation to estimate substitution patterns.

For four electronics categories, we collect standard attributes from product detail pages to compare our approach with attribute-based mixed logit: 18 attributes for "Tablets," 13 for "Monitors," 20 for "Memory Cards," and 14 for "Headphones" (see Appendix F for a full list).

4.2 | Estimation Results

We first compare our approach to benchmark models in the four electronics categories for which we collected attribute data. Since we have a relatively large number of attributes in these categories, we reduce their dimensionality via PCA and apply Algorithm 1 for

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 | Predicted second choice probabilities and their data counterparts. Letters in parentheses indicate book genres: F = Fantasy, M = Mystery, and S = Self-Help.

Panel A. Predicted second-choice probabilities when first choice is *Dopamine Detox (S)*

Experimental data		Plain logit		Attribute-based mixed logit		Review-based mixed logit	
Book	**Prob.**	**Book**	**Prob.**	**Book**	**Prob.**	**Book**	**Prob.**
Don't Believe (S)	0.353	Please Tell Me (M)	0.171	Don't Believe (S)	0.220	Don't Believe (S)	0.266
Art of Letting Go (S)	0.249	The Inmate (M)	0.147	Please Tell Me (M)	0.153	Art of Letting Go (S)	0.169
Please Tell Me (M)	0.112	Don't Believe (S)	0.143	Art of Letting Go (S)	0.142	Please Tell Me (M)	0.134
The Inmate (M)	0.094	The Housemaid (M)	0.139	The Housemaid (M)	0.134	The Inmate (M)	0.123
The Housemaid (M)	0.057	Art of Letting Go (S)	0.106	The Inmate (M)	0.131	The Housemaid (M)	0.101
Serpent & Wings (F)	0.042	Court of Ravens (F)	0.096	Court of Ravens (F)	0.101	Court of Ravens (F)	0.070
Court of Ravens (F)	0.034	Serpent & Wings (F)	0.088	Serpent & Wings (F)	0.060	Serpent & Wings (F)	0.057
The Ritual (M)	0.031	The Ritual (M)	0.057	The Ritual (M)	0.031	The Ritual (M)	0.043
Ashes & Star (F)	0.030	Ashes & Star (F)	0.054	Ashes & Star (F)	0.027	Ashes & Star (F)	0.037

Panel B. Predicted second-choice probabilities when first choice is *Please Tell Me (M)*

Experimental data		Plain logit		Attribute-based mixed logit		Review-based mixed logit	
Book	**Prob.**	**Book**	**Prob.**	**Book**	**Prob.**	**Book**	**Prob.**
The Inmate (M)	0.325	The Inmate (M)	0.153	The Inmate (M)	0.162	The Inmate (M)	0.173
The Housemaid (M)	0.250	Don't Believe (S)	0.149	The Housemaid (M)	0.151	The Housemaid (M)	0.169
Don't Believe (S)	0.088	The Housemaid (M)	0.145	Don't Believe (S)	0.136	Don't Believe (S)	0.115
Art of Letting Go (S)	0.066	Dopamine Detox (S)	0.137	Dopamine Detox (S)	0.115	Court of Ravens (F)	0.107
Serpent & Wings (F)	0.066	Art of Letting Go (S)	0.110	Art of Letting Go (S)	0.106	Serpent & Wings (F)	0.107
Dopamine Detox (S)	0.063	Court of Ravens (F)	0.100	Court of Ravens (F)	0.102	Dopamine Detox (S)	0.101
Court of Ravens (F)	0.058	Serpent & Wings (F)	0.092	Serpent & Wings (F)	0.100	Art of Letting Go (S)	0.095
The Ritual (M)	0.056	The Ritual (M)	0.059	The Ritual (M)	0.064	The Ritual (M)	0.068
Ashes & Star (F)	0.029	Ashes & Star (F)	0.056	Ashes & Star (F)	0.064	Ashes & Star (F)	0.063

Panel C. Predicted second-choice probabilities when first choice is *Ashes & Star (F)*

Experimental data		Plain logit		Attribute-based mixed logit		Review-based mixed logit	
Book	**Prob.**	**Book**	**Prob.**	**Book**	**Prob.**	**Book**	**Prob.**
Serpent & Wings (F)	0.275	Please Tell Me (M)	0.159	Please Tell Me (M)	0.184	Please Tell Me (M)	0.176
Court of Ravens (F)	0.243	The Inmate (M)	0.136	The Inmate (M)	0.158	The Housemaid (M)	0.152
The Inmate (M)	0.105	Don't Believe (S)	0.133	The Housemaid (M)	0.138	The Inmate (M)	0.150
The Ritual (M)	0.082	The Housemaid (M)	0.129	Serpent & Wings (F)	0.123	Court of Ravens (F)	0.117
Please Tell Me (M)	0.080	Dopamine Detox (S)	0.122	The Ritual (M)	0.097	Serpent & Wings (F)	0.104
The Housemaid (M)	0.070	Art of Letting Go (S)	0.098	Court of Ravens (F)	0.086	Don't Believe (S)	0.086
Don't Believe (S)	0.052	Court of Ravens (F)	0.089	Don't Believe (S)	0.081	Dopamine Detox (S)	0.080
Art of Letting Go (S)	0.048	Serpent & Wings (F)	0.082	Art of Letting Go (S)	0.071	Art of Letting Go (S)	0.072
Dopamine Detox (S)	0.045	The Ritual (M)	0.053	Dopamine Detox (S)	0.061	The Ritual (M)	0.063

model selection.[18] As in Section 3, we let the candidate set include price and the first $P = 6$ principal components.

Table 3 shows the results. In all four categories, unstructured data meaningfully improves model fit relative to both plain logit and attribute-based mixed logit, suggesting that our approach captures information about substitution patterns beyond that contained in standard attributes. This finding is particularly noteworthy given that one might expect technical specifications of electronics products—captured by our extensive list of attributes—to be highly relevant for consumer choices.

These performance differences translate into different substitution patterns. For example, consider diversion ratios for the

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

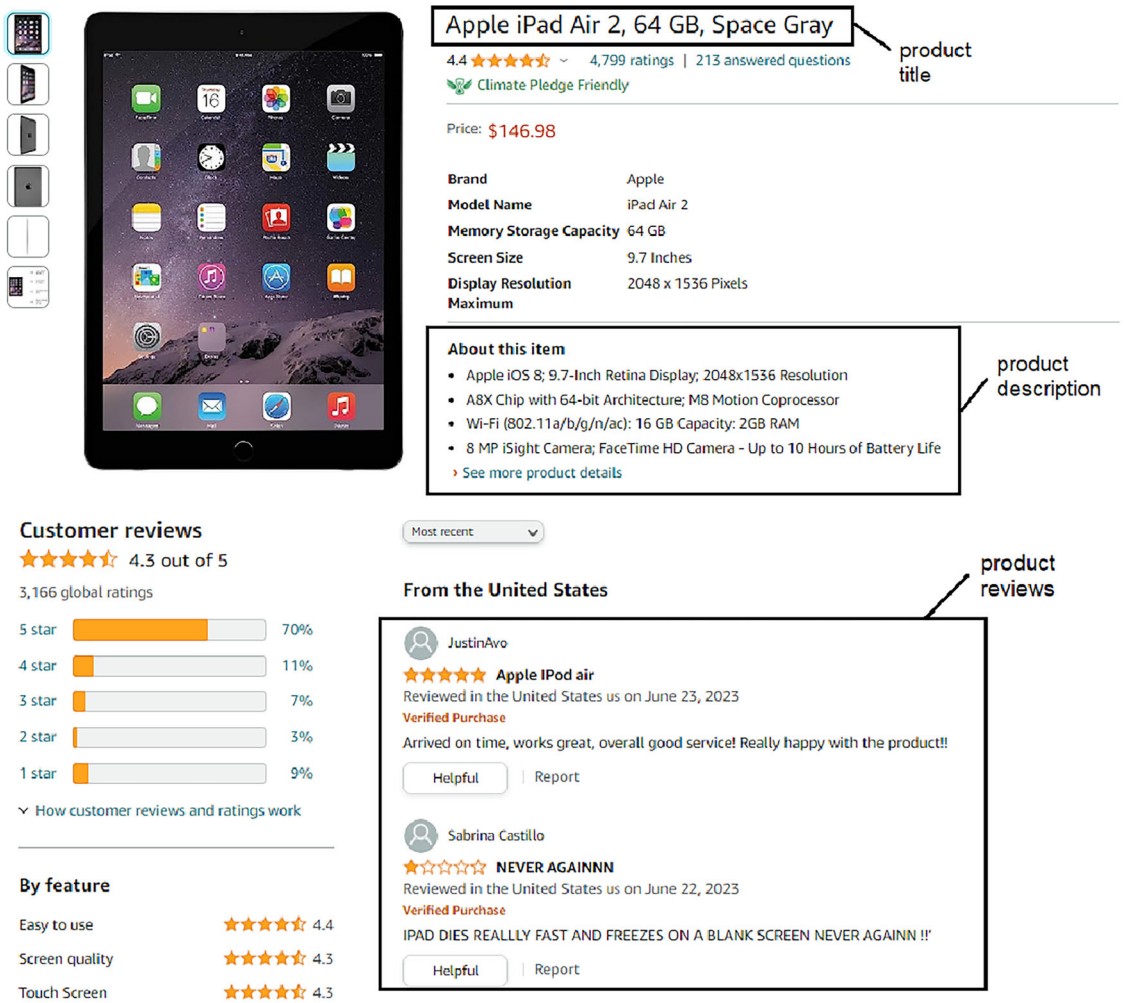


FIGURE 5 | Example: image and text data collected from product detail pages.

TABLE 3 | *AIC* comparison of our approach and attribute-based mixed logit. The table reports the *AIC* of the specification selected by Algorithm 1 within each class of demand models.

	AIC	**Δ*AIC* relative to plain logit**
Category: Tablets		
Mixed logit with attributes	4293.2	−54.9
Mixed logit with unstructured data	4236.6	−111.5
Category: Monitors		
Mixed logit with attributes	1376.4	−12.3
Mixed logit with unstructured data	1366.4	−22.4
Category: Memory cards		
Mixed logit with attributes	2709.3	−72.6
Mixed logit with unstructured data	2685.1	−96.8
Category: Headphones		
Mixed logit with attributes	9882.0	−4.5
Mixed logit with unstructured data	9875.8	−10.7

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

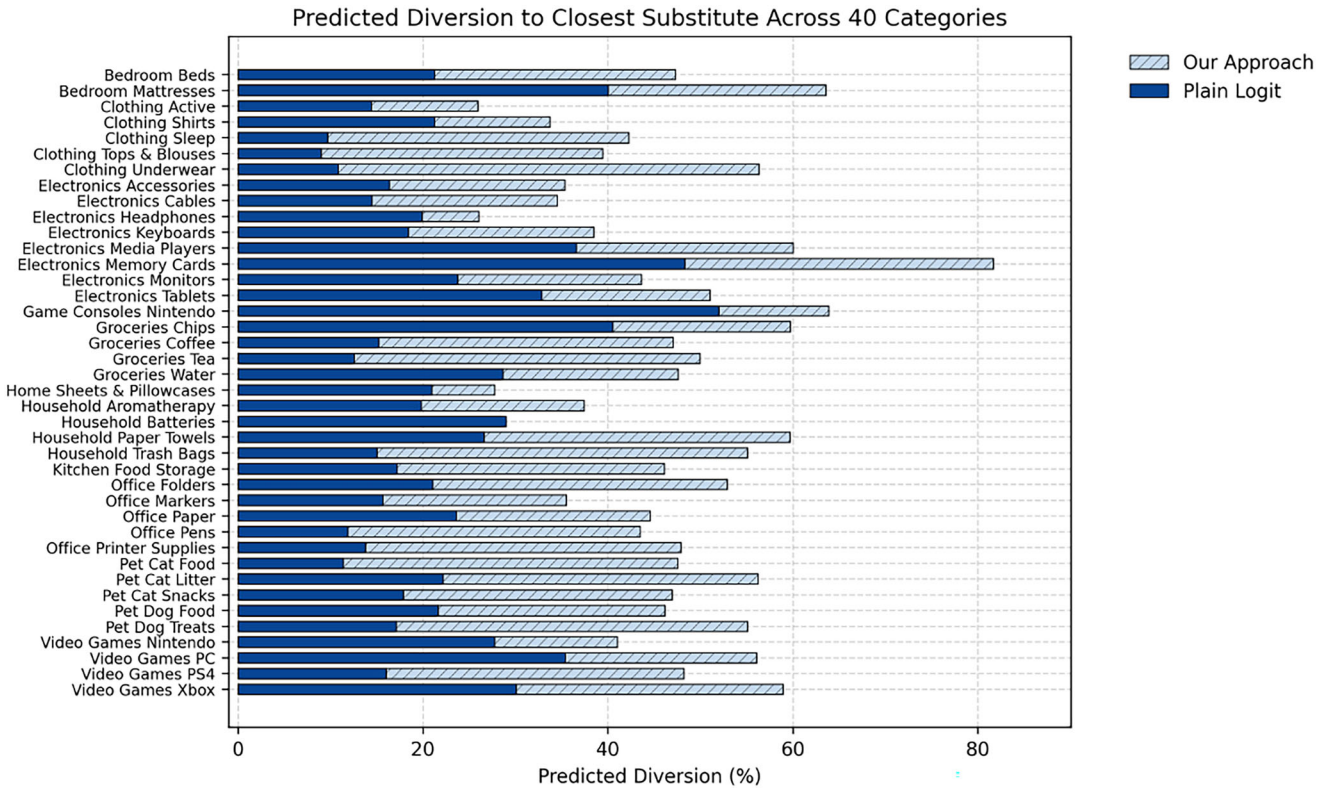


FIGURE 6 | Estimated diversion ratios to closest substitutes.

Note: This figure plots the estimated diversion ratios to the closest substitutes, $\max_k \hat{s}_{j\to k}$, averaged across products j in each category.

category of tablets (see Tables A2– A4). Relative to the plain logit, the selected *Description ST* model produces more intuitive substitution patterns: kids' tablets are close substitutes (Fire Kids and Fire HD8 Kids), iPads are close substitutes (iPad 9.7 and iPad 10.2), and so on. The predicted diversions in this model also differ markedly from those in the mixed logit with attributes. However, in the absence of second choice data, we cannot directly test how well different models predict diversions counterfactually, as we do in Section 3. We therefore rely on the fit comparisons reported in Table 3, which favor the *Description ST* model.

Next, we expand the analysis to all 40 categories. Because observed attributes are not available in all of them, we benchmark results against the plain logit model (see Table A5 and Figure A6). As a rule of thumb, we consider a model to have strong support over a simpler model if it lowers *AIC* by at least 2.0, and very strong support if it lowers *AIC* by at least 5.0.[19] We find that our approach reduces *AIC* by at least 5.6 in every category. On average, the reduction is 23.3, reaching 111.5 in some categories.[20] These results suggest that our approach consistently captures signals of substitution from unstructured data across a wide range of categories.

A well-known limitation of the plain logit is that diversion ratios depend only on market shares rather than attribute similarity, leading to overly flat diversions (Conlon et al. 2023). Figure 6 plots the estimated diversion ratio to the closest substitute, $\max_k \hat{s}_{j\to k}$, averaged across products j. The plain logit yields relatively small diversions, about 22% on average, while our approach increases this average to 47% and, in some categories, to as high as 60-80%. These results indicate that our approach produces substantially more variable diversion ratios, suggesting it better identifies close substitutes.

4.3 | Relevance of Different Data Types

Lastly, we analyze which types of unstructured data yield the largest fit improvements. It would be misleading to report only the best-performing model in each product category, as multiple text or image models may achieve similar *AIC* improvements over plain logit. We therefore follow the statistical literature on model selection and report Akaike weights (Burnham and Anderson 2004). Formally, let d denote data type (images, titles, descriptions, or reviews), and let AIC_d denote the lowest *AIC* among all specifications that use data type d. For each data type d, we compute the Akaike weight w_d as

$$w_d = \frac{\exp(-\Delta_d/2)}{\sum_k \exp(-\Delta_k/2)} \tag{3}$$

where the summation in the denominator is over all four data types, $\Delta_d = AIC_d - AIC_{min}$, and $AIC_{min} = \min_d AIC_d$ is the lowest *AIC* across *all* four data types. We interpret w_d heuristically as the posterior probability that the mixed logit model based on data type d is the best model given the data.[21]

Figure 7 shows the estimated Akaike weights by category and data type, revealing substantial variation across categories. Many results are difficult to anticipate *a priori*. For example, although

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

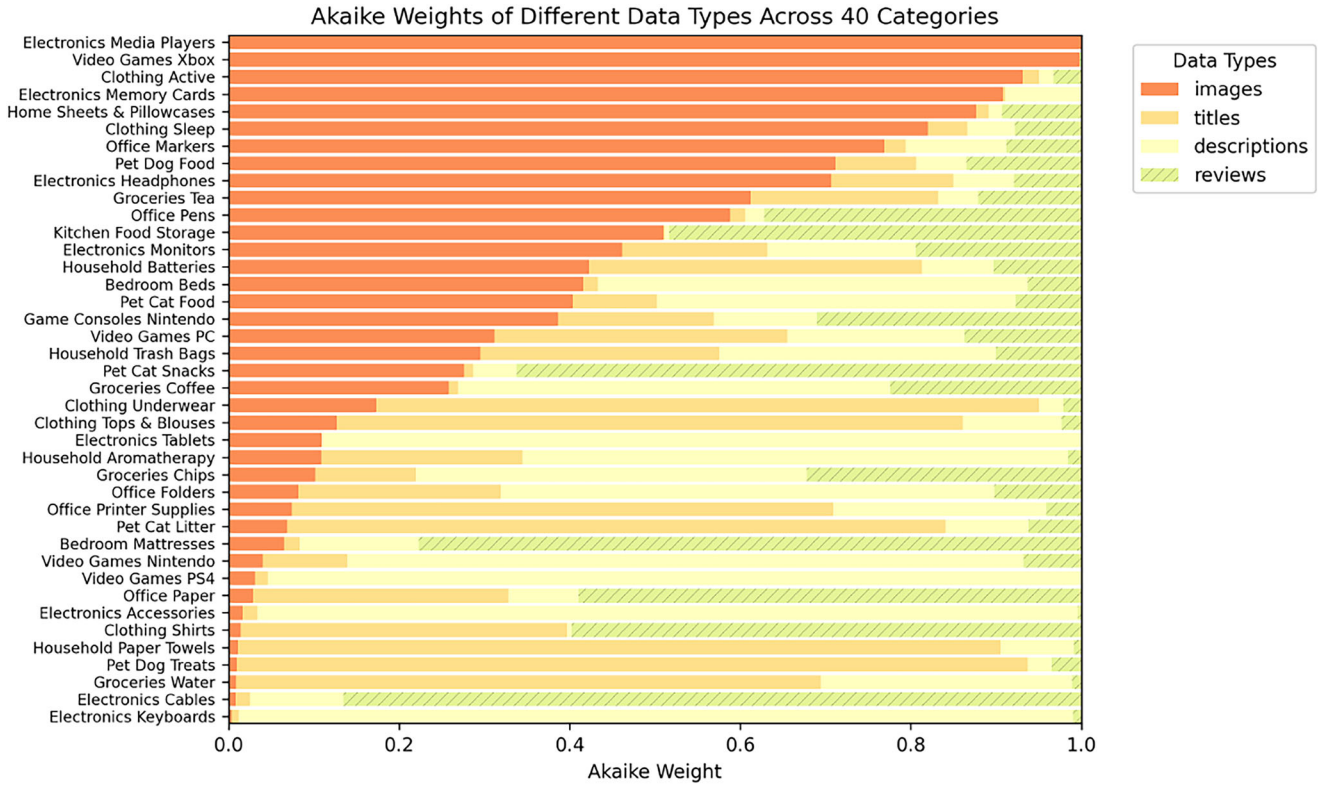


FIGURE 7 | Akaike weights of different types of unstructured data across 40 product categories.

Note: These weights reflect the relative performance of data types in capturing substitution patterns from choice data. We compute Akaike weights using the formula in (3).

one might expect visual features to be most relevant for clothing, only two of the five clothing categories ("Activewear" and "Sleep") place more weight on images than text. In "Tops & Blouses" and "Underwear," product titles are most predictive of substitution, while in "Shirts," reviews perform best. Similarly, we may expect descriptions and reviews of video games to be most informative, analogous to books. In practice, we find that images contain substantially more information about substitution in the "Video Games Xbox" category.

These results highlight that researchers cannot reliably predict which data types best capture substitution. In practice, it is therefore important to collect multiple data types and use model selection to choose the one that best captures substitution.

5 | A Practitioner's Guide for Using Text and Images

In this section, we provide guidance for practitioners interested in applying our method. We highlight several practical lessons from Sections 3 and 4, clarify which counterfactuals our approach can and cannot accommodate, and suggest promising directions for future research.

5.1 | Data Choice and Model Selection

The main takeaway from our results is that text and image data contain valuable information about substitution patterns, making these unstructured data useful for demand estimation. At the same time, our multi-category analysis in Section 4 shows that researchers may not be able to predict in advance which data type will best capture substitution in a given category. We therefore recommend collecting various types of unstructured data and performing model selection as in Sections 3 and 4. Algorithm 1 in Section 2 provides a practical heuristic for model selection: in our experiment, it successfully identifies the specification that produces the best counterfactual second choice predictions.

A natural question is whether researchers should add observed attributes to our approach when they are available. In our application, adding observed attributes to the selected review-based model does not improve fit (see Section 3.3). One explanation is that unstructured data may already encode the choice-relevant attributes, rendering structured attributes redundant. Nevertheless, in other applications, researchers may still want to test whether including observed attributes helps estimate substitution patterns. For example, researchers can expand the set of candidate variables in Algorithm 1 to include the observed attributes themselves or, when many attributes are available, their principal components.

5.2 | Counterfactual Analysis and Its Boundaries

Our approach is well suited for applications that assess consumer responses to price changes while holding embeddings

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

fixed, including optimal pricing, merger simulations, corrective taxes, and markup estimation (Berry and Haile 2024). In these applications, embeddings may be correlated with prices in the observed data, but they, and hence the demand system, should be held fixed in counterfactual simulations. Under this assumption, researchers can compute counterfactual market shares, prices, and welfare using standard tools from demand analysis. For instance, consumer welfare can be computed as the area under the demand curve (Small and Rosen 1981). The fact that embeddings have no direct interpretation does not preclude these analyses.

Our approach can also predict responses to changes in product availability, for example, to study firms' assortment choices or consumer surplus from new products (Petrin 2002). If new product launches are observed in the data, researchers can apply our method directly, since our results show it performs well at predicting consumer responses to product removals. On the other hand, when evaluating a hypothetical new product, researchers must construct its embeddings from pre-launch information, such as images and descriptions from marketing materials or early retailer listings.

By contrast, our approach is less suitable in applications where embeddings may change in counterfactuals. For example, if customer reviews mention value relative to price (e.g., *"Great quality for the price!"* or *"Good product, but overpriced."*), embeddings constructed from them may shift when prices change. Researchers can diagnose this issue by checking how frequently price-related language appears in reviews and, if necessary, modeling how price changes affect embeddings. Another example is simulations of mergers that may induce firms to redesign products, which would alter embeddings and hence substitution patterns. Studying such mergers is not possible with our approach unless researchers impose additional assumptions on firm conduct and on how embeddings map into attributes that firms can control.

5.3 | Future Research Directions

Our paper shows that text and image data contain valuable and easily extractable information for estimating substitution patterns. This finding opens several promising directions for future research.

First, newer ML models might extract substitution patterns from texts and images more effectively. For example, Qwen3, OpenAI, and Gemini text embeddings perform consistently well across diverse natural language tasks and may thus generalize to demand estimation (Lee et al. 2025; Zhang et al. 2025). Because our method is modular, researchers can easily try alternative embeddings and test whether that improves performance. In addition, researchers can fine-tune existing models to construct embeddings optimized for counterfactual predictions. An open question is how much fine-tuning improves upon pre-trained models and whether these gains come without excessive computational costs. To facilitate future research on alternative embedding models, we make our experimental dataset and estimation code publicly available.[22]

Second, our method incorporates embeddings into the demand model as if they were the true attributes driving substitution, when in practice embeddings may be only imperfect proxies of these attributes. Our results show that, despite this limitation, embeddings are sufficiently informative to improve counterfactual predictions. Nevertheless, future research should explore whether explicitly accounting for error in these proxies can further improve performance. For an early contribution in this area, see Christensen and Compiani (2026).

Third, other demand models may leverage unstructured data more effectively. In an earlier version of this paper, we tested the pairwise combinatorial logit model of Koppelman and Wen (2000), which allows pairwise utility correlations to depend on distances in the embedding space. That model performed worse than mixed logit, not only when using unstructured data but also when using distances in observed attributes, suggesting that functional form assumptions can significantly impact counterfactual performance. A fruitful direction for future research might be to combine unstructured data with more flexible demand models, such as that in Compiani (2022).

Finally, because our focus is on recovering substitution patterns, we largely abstract away from price endogeneity. Endogeneity is not a concern in our experiment because we randomize prices. In the e-commerce application, we rely on product fixed effects to capture unobserved attributes (e.g., quality) that may correlate with prices. More generally, prices may correlate with unobserved demand shocks that vary across markets and over time. To address this, one could combine our approach with existing methods for handling price endogeneity using instrumental variables (Goolsbee and Petrin 2004; Berry et al. 2004).[23]

6 | Conclusion

In this paper, we show how researchers can incorporate unstructured text and image data in demand estimation to recover substitution patterns. Our approach extracts low-dimensional features from product images and textual descriptions, integrating them into a standard mixed logit model with random coefficients. Using experimental data, we show that our approach outperforms standard attribute-based models in counterfactual predictions of second choices. We further validate our method with e-commerce data across dozens of categories and find that text and image data consistently help identify close substitutes within each category.

Acknowledgments

We thank Steve Berry, Alex Burnap, Matt Gentzkow, Philip Haile, Yufeng Huang, Joonhwi Joo, Jon McClure, Carl Mela, Sanjog Misra, Olivia Natan, Matt Osborne, Katja Seim, Adam Smith, Yuyan Wang, Kevin Williams and seminar participants at Bocconi University, Chicago Booth, London Business School, Stockholm School of Economics, University of Cologne, Nova SBE, Marketing Science Conference 2022, QME Conference 2025, SICS Conference 2024, AI/ML Yale Conference, and SED Conference for helpful comments. We also thank Celina Park, Tanner Parsons, James Ryan, Janani Sekar, Andrew Sharng, Franklin She, Adam Sy, and Vitalii Tubdenov for outstanding research assistance. Giovanni Compiani gratefully acknowledges support from the Booth Center for Applied AI.

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Endnotes

[1] This criticism of attribute-based models has a long history in the economics literature. For example, Hausman (1994, 229) skeptically remarks that applying such models to French champagne choices would require researchers to somehow quantify the bubble content.

[2] The package is available on PyPI and in the public GitHub repository: https://github.com/deep-logit-demand/deeplogit. Our package heavily borrows from the existing package *Xlogit* for GPU-accelerated estimation of mixed logit models, developed by Arteaga et al. (2022).

[3] We refer to second choice predictions as "counterfactual predictions" in the sense that second choices are not used during estimation, which is based solely on first choices.

[4] Berry et al. (2004) show that second choice data are useful for estimating substitution patterns, and Conlon et al. (2023) estimate demand from aggregate data on first-choice probabilities and a subset of second-choice probabilities. Conceptually, our validation approach reverses the logic of these papers: we only use first-choice data for estimation and reserve second choices for model validation. In a similar spirit to our approach, Raval et al. (2022) use hospital closures induced by natural disasters to compare the ability of various models to predict diversions.

[5] Döpper et al. (2024) remark that estimating demand across multiple categories "would be difficult to implement at scale because it would require category by-category assessments about which characteristics are appropriate to include and whether or not relevant data are available." Our approach can address this problem if researchers have access to unstructured data.

[6] These models were originally trained to classify images into labeled categories (e.g., "cup," "book," or "sofa"). However, since our goal is not to label products but to measure visual features that distinguish them from each other, we remove the classification layer from these models and work directly with embeddings.

[7] TF-IDF assigns more weight to words that are frequent in a specific text but rare across others, thus emphasizing unique words. This makes it more likely that text embeddings capture distinctive words that differentiate products from one another.

[8] The model trained by Reimers and Gurevych (2019) is a more efficient version of the widely used BERT network (Devlin et al. 2018).

[9] While we could select the specification that best matches observed second choices in our experiment, this strategy is unavailable to most researchers who usually only have first-choice data. We thus propose a model selection algorithm that relies solely on first choices.

[10] We collect the first five reviews displayed on each product detail page at the time of data collection.

[11] Prices were drawn from a discrete uniform distribution ranging from \$3 to \$7.

[12] The pre-registration document can be accessed at https://tinyurl.com/ekjp4pfp. At the time of pre-registration, we planned to estimate a different choice model—the pairwise combinatorial logit (PCL) model from Koppelman and Wen (2000)—which generalizes the logit model by allowing utility correlations across product pairs. We switched to a mixed logit model at a later stage because we found it to perform better. We adhered to our pre-registered protocol for all key aspects of survey design, including sample size, sample selection, and choice task construction.

[13] We do not include a random coefficient on rank_{ij} in any model specification because our focus is on estimating substitution patterns using observed characteristics or principal components.

[14] Since the principal components do not vary over time, the mean of θ_i is absorbed by the product fixed effects δ_j. The same holds for the mean of β_i in the mixed logit model with attributes. In addition, we include the mean price coefficient $\bar{\alpha}$ in all specifications, regardless of whether the random coefficient on price is selected by the model selection algorithm.

[15] The first six principal components together explain 70%–80% of the variance in embeddings (Figure A5).

[16] We chose this approach because including all principal components from unstructured data and all observed attributes in the candidate set for Algorithm 1 would be too computationally burdensome.

[17] We do not observe product rankings, so we do not include them in our demand models.

[18] We normalize each attribute to have mean zero and variance one before applying PCA to prevent attributes with larger variance or units from dominating the principal components.

[19] When a simpler model is nested within a more complex one with only one additional parameter, applying these AIC thresholds is approximately equivalent to conducting a likelihood ratio test at the 5% and 1% significance levels. In this case, the reduction in AIC is given by $\Delta AIC = 2 - \lambda_{LR}$, where λ_{LR} is the likelihood ratio statistic, meaning the ΔAIC thresholds of 2.0 and 5.0 correspond to the likelihood ratio thresholds 4.0 and 7.0 (p-values 0.0455 and 0.008).

[20] In some categories, our estimates of the price coefficient α are positive, likely due to correlation between prices and unobserved demand shocks, even after accounting for product fixed effects. While instrumental variables could address this, we do not pursue this approach, as addressing price endogeneity across many product categories is orthogonal to the contribution of our paper. Instead, we focus on counterfactuals, such as the diversion ratios from removing a product, which remain valid even if the price coefficient is positive.

[21] More precisely, in large samples, w_d reflects the probability that this class of models is, in fact, the best model for the data in the sense of Kullback-Leibler information (Burnham and Anderson 2004, 272).

[22] Replication codes and experimental data are available in our public repository: https://github.com/ilyamorozov/DeepLogitReplication. Separately, the Python package for implementing our method is available on PyPI and at: https://github.com/deep-logit-demand/deeplogit.

[23] For example, one could estimate product-market fixed effects and random coefficient variances via maximum likelihood, and then estimate a two-stage least squares regression of the estimated fixed effects on prices and principal components. The principal components of competing products could serve as excluded instruments for prices.

References

Akaike, H. 1998. "Information Theory and an Extension of the Maximum Likelihood Principle." In *Selected Papers of Hirotugu Akaike*, 199–213. Springer.

Allcott, H., B. B. Lockwood, and D. Taubinsky. 2019. "Regressive Sin Taxes, With An Application to the Optimal Soda Tax." *The Quarterly Journal of Economics* 134: 1557–1626.

Arteaga, C., J. Park, P. B. Beeramoole, and A. Paz. 2022. "xlogit: An Open-Source Python Package for GPU-Accelerated Estimation of Mixed Logit Models." *Journal of Choice Modelling* 42: 100339.

Backus, M., C. Conlon, and M. Sinkinson. 2021. "Common Ownership and Competition in the Ready-To-Eat Cereal Industry." Working Paper.

Battaglia, L., T. Christensen, S. Hansen, and S. Sacher. 2024. "Inference for Regression with Variables Generated by AI or Machine Learning." arXiv:2402.15585.

Berry, S., J. Levinsohn, and A. Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63: 841–890.

Berry, S., J. Levinsohn, and A. Pakes. 1999. "Voluntary Export Restraints on Automobiles: Evaluating A Trade Policy." *American Economic Review* 89: 400–431.

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Berry, S., J. Levinsohn, and A. Pakes. 2004. "Differentiated Products Demand System from a Combination of Micro and Macro Data: The New Car Market." *Journal of Political Economy* 112: 68–105.

Berry, S. T., and P. A. Haile. 2024. "Nonparametric Identification of Differentiated Products Demand Using Micro Data." *Econometrica* 92: 1135–1162.

Burnham, K. P., and D. R. Anderson. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods & Research* 33: 261–304.

Cer, D., M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. 2017. "Semeval-2017 Task 1: Semantic Textual Similarity-Multilingual and Cross-Lingual Focused Evaluation." arXiv preprint arXiv:1708.00055.

Cer, D., Y. Yang, S. yi Kong, et al. 2018. "Universal Sentence Encoder." arXiv:1803.11175.

Chollet, F. 2017. "Xception: Deep Learning With Depthwise Separable Convolutions." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258. IEEE.

Christensen, T., and G. Compiani. 2026. "From Unstructured Data to Demand Counterfactuals: Theory and Practice." arXiv preprint, arXiv:2601.05374.

Compiani, G. 2022. "Market Counterfactuals and the Specification of Multiproduct Demand: A Nonparametric Approach." *Quantitative Economics* 13: 545–591.

Conlon, C., J. Mortimer, and P. Sarkis. 2023. "Estimating Preferences and Substitution Patterns from Second Choice Data Alone." Tech. Rep., Working Paper.

Conlon, C., and J. H. Mortimer. 2021. "Empirical Properties of Diversion Ratios." *The RAND Journal of Economics* 52: 693–726.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805.

Dew, R. 2024. "Adaptive Preference Measurement with Unstructured Data."*Management Science*.

Döpper, H., A. MacKay, N. Miller, and J. Stiebale. 2024. "Rising Markups and the Role of Consumer Preferences." Tech. Rep., National Bureau of Economic Research.

Dotson, J. P., M. A. Beltramo, E. M. Feit, and R. C. Smith. 2019. "Modeling the Effect of Images on Product Choices." Working Paper.

Federal Trade Commission. 2022. "Demand System Estimation and Its Application to Horizontal Merger Analysis." Tech. Rep., Federal Trade Commission.

K Goldberg, P. 1995. "Product Differentiation and Oligopoly in International Markets: The Case of the US Automobile Industry." *Econometrica: Journal of the Econometric Society* 63, no. 4: 891–951.

Goodfellow, I., Y. Bengio, A. Courville, and Y. Bengio. 2016. *Deep Learning* Vol. 1. MIT Press.

Goolsbee, A., and A. Petrin. 2004. "The Consumer Gains From Direct Broadcast Satellites and the Competition with Cable TV." *Econometrica* 72: 351–381.

Greminger, R., Y. Huang, and I. Morozov. 2023. "Make Every Second Count: Time Allocation in Online Shopping." Working Paper.

Han, S., and K. Lee. 2025. "Copyright and Competition: Estimating Supply and Demand with Unstructured Data." arXiv preprint, arXiv:2501.16120.

Hausman, J. A. 1994. *Valuation of New Goods Under Perfect and Imperfect competition*. National Bureau of Economic Research.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. IEEE.

Hoch, S. J., B.-D. Kim, A. L. Montgomery, and P. E. Rossi. 1995. "Determinants of Store-Level Price Elasticity." *Journal of Marketing Research* 32: 17–29.

Koppelman, F. S., and C.-H. Wen. 2000. "The Paired Combinatorial Logit Model: Properties, Estimation and Application." *Transportation Research Part B: Methodological* 34: 75–89.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks."*Advances in Neural Information Processing Systems* 25.

Lee, J., F. Chen, S. Dua, et al. 2025. "Gemini Embedding: Generalizable Embeddings From Gemini." arXiv preprint, arXiv:2503.07891.

Lee, K. 2024. "Generative Brand Choice." PhD diss., University of Chicago.

Magnolfi, L., J. McClure, and A. Sorensen. 2022. "Triplet Embeddings for Demand Estimation." Working Paper.

McFadden, D., and K. Train. 2000. "Mixed MNL Models for Discrete Response." *Journal of Applied Econometrics* 15: 447–470.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." arXiv preprint, arXiv:1301.3781.

Nevo, A. 2000. "Mergers with Differentiated Products: The Case of the Ready-To-Eat Cereal Industry."*The RAND Journal of Economics* 395–421.

Nevo, A. 2001. "Measuring Market Power in the Ready-To-Eat Cereal Industry." *Econometrica* 69: 307–342.

Petrin, A. 2002. "Quantifying the Benefits of New Products: The Case of the Minivan." *Journal of Political Economy* 110: 705–729.

Quan, T. W., and K. R. Williams. 2019. *Extracting Characteristics From Product Images and its Application to Demand Estimation*. University of Georgia, Department of Economics.

Raval, D., T. Rosenbaum, and N. E. Wilson. 2022. "Using Disaster-Induced Closures To Evaluate Discrete Choice Models of Hospital Demand." *The RAND Journal of Economics* 53: 561–589.

Reimers, N., and I. Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Russakovsky, O., J. Deng, H. Su, et al. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115: 211–252.

Seiler, S., A. Tuchman, and S. Yao. 2021. "The Impact of Soda Taxes: Pass-Through, Tax Avoidance, and Nutritional Effects." *Journal of Marketing Research* 58: 22–49.

Simonyan, K., and A. Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." in International Conference on Learning Representations.

Sisodia, A., A. Burnap, and V. Kumar. 2024. "Generative Interpretable Visual Design: Using Disentanglement for Visual Conjoint Analysis." *Journal of Marketing Research* 62, no. 3: 405–428.

Small, K. A., and H. S. Rosen. 1981. "Applied Welfare Economics With Discrete Choice Models." *Econometrica: Journal of the Econometric Society* 49, no. 1: 105–130.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. "Rethinking the Inception Architecture for Computer Vision." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826. IEEE.

Vilcassim, N. J., and P. K. Chintagunta. 1995. "Investigating Retailer Product Category Pricing from Household Scanner Panel Data." *Journal of Retailing* 71: 103–128.

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Zhang, Y., M. Li, D. Long, et al. 2025. "Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models." arXiv preprint, arXiv:2506.05176.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Data S1 Data S2

17562171, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1756-2171.70052 by Imperial College London, Wiley Online Library on [30/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License