

Demand Estimation with Text and Image Data^{*}

Giovanni Compiani

University of Chicago

Ilya Morozov

Northwestern University

Stephan Seiler

Imperial College London
& CEPR

We propose a demand estimation method 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 outperforms standard attribute-based models at counterfactual predictions of second choices. We also apply it to 40 product categories on Amazon.com and consistently find that unstructured data identifies close substitutes within each category.

Keywords: Demand Estimation, Unstructured Data, Deep Learning

^{*}We thank Steve Berry, Alex Burnap, Matt Gentzkow, Joonhwi Joo, Philip Haile, Yufeng Huang, 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, 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. Contact information: Giovanni Compiani (gio1compiani@gmail.com), Ilya Morozov (ilya.morozov@kellogg.northwestern.edu), Stephan Seiler (stephan.a.seiler@gmail.com). Giovanni Compiani gratefully acknowledges support from the Center for Applied AI at Booth.

1 Introduction

Many problems in economics and marketing—such as merger analysis, tariff evaluation, and optimal pricing—require researchers to estimate demand for differentiated products. A standard approach to these problems 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 attributes that differentiate products. Instead, they rely on third-party data, where attributes are chosen based on unknown criteria, or gather their own data, subjectively selecting which attributes to collect. The collected attributes may not align with those most relevant to consumer choices. 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 text and image data in demand estimation to recover substitution patterns. We propose an approach that uses product images, titles, descriptions, and customer reviews—unstructured data that are widely available even in markets where collecting product attributes is challenging. Using pre-trained deep learning models, we transform images and texts into vector representations—*embeddings*. We then apply Principal Component Analysis (PCA) to reduce dimensionality and capture the main dimensions of product differentiation. We incorporate these principal components into a mixed logit demand model (Berry et al., 1995; Nevo, 2001), interacting them with random coefficients similar to how researchers usually treat observed product attributes in demand estimation.

This approach enables researchers to incorporate hard-to-quantify product attributes, such as visual design from images and functional benefits from text, while also circumventing the need to subjectively choose which observed attributes to collect. Given the prevalence of unstructured data, we see our approach as a valuable addition to the toolbox of empirical researchers. To make it easier for others to apply our approach, we provide a publicly available Python package, *DeepLogit*, alongside this paper.²

To show that the proposed method performs well at recovering substitution patterns, we validate it extensively on several datasets. Validation in demand estimation is challenging because substitution patterns are usually unobserved, leaving no clear “ground truth” to evaluate model predictions. To overcome this challenge, we design a choice ex-

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

²The package is available on PyPI and in the public GitHub repository: 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).

periment where participants choose a book from a list of options. We randomize both prices and rankings to identify substitution patterns. Crucially, we elicit both first and second choices, estimate different demand models using only first choices, and evaluate their performance based on how well they predict second choices counterfactually.³ Conditional second choices are informative about substitution because they reveal which product a consumer would substitute to if their preferred option was unavailable. Recording second choices thus enables us to assess how well each demand model predicts substitution patterns.⁴

Using these experimental data, we show that both text and images contain useful information about substitution patterns. Intuitively, images capture substitution because book covers visually convey information about genre. Text performs even better, likely because descriptions and reviews capture nuanced details about book plots, helping identify similar books even within the same genre. We then show that the best performing model—based on customer reviews—improves counterfactual predictions of second choices relative to the standard mixed logit with observed attributes. In fact, the improvement relative to that model is on par with the improvement that the mixed logit achieves over the plain logit model without random coefficients. Thus, in this application, we can recover substitution from unstructured data alone, and in fact can achieve better performance than is possible even with attribute data.

To assess the economic importance of these improvements, we simulate hypothetical mergers of book publishers. We find that differences in estimated substitution patterns are large enough such that an antitrust agency relying on the standard mixed logit with attributes might incorrectly approve mergers that should be challenged and block mergers that should be approved. Thus, our results show that incorporating text and image data yields improvements in estimated substitution patterns that are both statistically and economically significant.

We also validate our approach on observational data. We apply it to 40 product categories on Amazon.com, including groceries, pet food, office supplies, beauty products, electronics, video games, and clothing. We combine data on online purchases from the Comscore Web Behavior panel with product images, titles, descriptions, and reviews collected from Amazon’s product detail pages. Here, our contribution is twofold. First, we

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

⁴Second choices are related to the notion of diversion ratios, which reflect the degree of substitutability and thus the intensity of price competition between two products (Shapiro, 1995; Conlon and Mortimer, 2021; Conlon et al., 2023). Recovering them is crucial for antitrust authorities trying to predict price increases from horizontal mergers and for managers optimizing assortments and prices.

show that text and images contain useful signals about substitution in all studied categories, highlighting that our approach is broadly applicable across markets. Second, we find—perhaps unexpectedly—that text performs best in some categories where images might be expected to be more useful, such as clothing, and vice versa. A practical takeaway is that researchers may benefit from collecting both text and images, even when only one seems relevant *a priori*.

Our paper contributes to the extensive literature on demand estimation. Specifically, 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 that incorporating the principal components of text and image embeddings into these models captures substitution patterns well. A key contribution of our paper is that we validate this approach extensively using both experimental and observational data.

This paper is also related to a vast literature in computer science that transforms unstructured data into lower-dimensional representations (Krizhevsky et al., 2012; Mikolov et al., 2013; Goodfellow et al., 2016). There is a growing literature on applications of pre-trained deep learning models in economics and marketing (see Battaglia et al., 2024 for a review). While most prior work focuses on predicting observed outcomes, our goal is fundamentally different: we aim to predict counterfactual quantities, such as substitution patterns. Predicting these quantities well is crucial for applications like merger simulations and optimal pricing. Yet, they are usually unobserved in choice data. As a result, we cannot apply standard cross-validation techniques and instead need to design a custom experiment that allows us to measure these otherwise unobserved quantities. To our knowledge, we are the first to use such an experiment for validation of demand models.⁵

Several other papers have proposed leveraging text and image data in demand estimation.⁶ Quan and Williams (2019) incorporate product image embeddings as shifters

⁵Berry et al. (2004) show that second choice data are useful for estimating substitution patterns, and Conlon et al. (2022) 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 treat second choices as “holdout” data for model validation. Additionally, Raval et al. (2022) use hospital closures induced by natural disasters to compare the ability of various models to predict diversions.

⁶Netzer et al. (2012) use data on the co-occurrence of products mentioned in online discussion forums to generate a visual representation of competing products, but they do not incorporate these measures into demand estimation. Dew (2024) uses image embeddings from pre-trained models to predict consumers’ ratings for products based on past ratings elicited in a survey. Sisodia et al. (2024) extract interpretable product attributes from images and use them to estimate preferences. Our approach is simpler and scales better across product categories, and thus, it may be preferable in applications where interpretability of extracted attributes is not essential.

of mean utilities. Similarly, [Lee \(2024\)](#) employs large-language models to predict utility intercepts of new products based on their textual descriptions. These methods predict mean utilities but are not designed to estimate utility covariances, which play a crucial role in shaping substitution.⁷ In contrast, our approach models utility covariances and can thus capture flexible substitution patterns—a key input for many empirical applications of interest. Further, while utility intercepts can be recovered from market shares, our goal is to capture substitution patterns that are unobserved. We measure them in an experiment by eliciting second choices, and we use these data for validation.⁸

A possible alternative to our use of unstructured data is to estimate substitution patterns using 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 random coefficients 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. Another complementary approach is proposed by [McClure \(2025\)](#), who shows how publicly available recommendations data from online platforms can also be used to estimate substitution patterns.

Our approach is well-suited for several empirical applications. Researchers can use it to address a wide range of economics and marketing questions that require accurate and flexible estimates of product substitution. This includes analyzing 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 altering assortments and prices. Additionally, researchers can apply this approach to study how multi-product firms optimize prices and promotions ([Hoch et al., 1995](#); [Vilcassim and Chintagunta, 1995](#)). Finally, because our approach circumvents the need to collect distinct attributes for each category, it may be valuable for researchers trying to estimate demand across many categories ([Döpper et al., 2024](#)).⁹

⁷For example, [Lee \(2024\)](#) inverts a plain logit model without random coefficients to extract product fixed effects for the prediction part of his analysis. Thus, his model restricts substitution patterns due to the IIA property.

⁸A recent paper by [Han and Lee \(2025\)](#) uses unstructured data to estimate substitution patterns among fonts in order to study the effects of copyright protection policies.

⁹[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

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.

Step 1. Extracting Embeddings from Texts and Images To extract information from product images, we extract their low-dimensional representations—*embeddings*—using pre-trained deep 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).¹⁰ The key advantage of using pre-trained models is that it reduces the computational burden of estimation while leveraging models trained on large-scale datasets, 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 customer 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.¹¹ While these count models are relatively simple, we view them as useful benchmarks because they can still detect attributes mentioned in titles, descriptions, and reviews. 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).¹² 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.

problem if researchers have access to unstructured data.

¹⁰These 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.

¹¹TF-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.

¹²The model trained by Reimers and Gurevych (2019) is a more efficient version of the widely used BERT network (Devlin et al., 2018).

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 used 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 therefore apply PCA to further reduce dimensionality (Backus et al., 2021).

The dimensionality reduction via PCA is attractive for multiple reasons. Raw embeddings are trained on general-purpose datasets to classify images into broad categories, such as “tablet” or “laptop.” However, to estimate demand, we need to analyze consumer choices *within a given category*. PCA helps us filter out the variation in embeddings necessary for sorting products into categories, allowing us to focus on dimensions most relevant for analyzing substitution within a category. Additionally, principal components are appealing because they are orthogonal to each other, which avoids multicollinearity issues and simplifies estimation of random coefficients in demand models.

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.¹³ We perform model selection using in-sample *AIC*. An advantage of using *AIC* is that it can be interpreted as a bias-corrected estimator of the expected relative Kullback-Leibler (KL) information based on the maximized log-likelihood function (Akaike, 1998). Further, in our experimental results, in-sample *AIC* strongly correlates with counterfactual performance, justifying its use for model selection (see Section 3.3 where we also implement *BIC* and cross-validation).

Our model selection algorithm considers a pool of candidate variables consisting of price and the first P principal components. In our application, we choose $P = 6$, which collectively account for 70-80% of variance in embeddings (see Appendix Figure A4). We first estimate all possible models with a random coefficient on a single variable and retain the best-fitting model. We then estimate all possible models with two random coefficients and assess whether fit improves relative to the best single-variable model. We continue to increase the number of variables K that have random coefficients until there is no further *AIC* improvement. We repeat this across all specifications—defined as a combination of

¹³While 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.

Algorithm 1: Model Selection

Input:

- Products $j = 1, \dots, J$
 - Specifications $m = 1, \dots, M$ (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)
-

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

a text or image model and data type (e.g., *USE Reviews*)—and choose the one with the lowest overall *AIC*. The full procedure is given in Algorithm 1.

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 ordering of principal components, which reflects the explained variance of embeddings but need not be predictive of preferences. For example, the approach might select a model that has random coefficients on the second and fourth principal components, but not the first or third.

Why Does This Approach Work? There are several reasons why texts and images may predict substitution. Consumers may choose based on how a product looks and how it is described by sellers or other consumers. Alternatively, even when consumers do not directly consider visual or textual descriptions, these descriptions may reflect attributes that drive 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 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 and uses it to capture substitution.

Given this interpretation, embeddings correlate with substitution patterns because they capture choice-relevant attributes, but the relationship need not be causal: changing product images or descriptions will not necessarily alter substitution patterns. Consequently, we cannot study questions like optimal product design or positioning unless we further unpack the link between embeddings and substitution, which is outside the scope of this paper. By contrast, our method is well-suited for counterfactuals where embeddings can be held fixed, such as optimal pricing or merger simulations. We provide a more detailed discussion of the counterfactuals enabled by our approach in Section 5.2.

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.

3 Validation with Experimental Data

We apply our method to data from a choice experiment to show it recovers substitution patterns more effectively than standard attribute-based methods. A key challenge is that we usually do not directly observe substitution patterns in the data, and thus do not have a “ground truth” to validate demand models. To address this, we design and implement a choice experiment that measures second-choice diversion ratios, which directly reflect substitution.

3.1 Experiment Design

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 from the list of ten options, we remove this book 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. 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 books’ rankings and prices across participants, keeping them fixed across choice tasks for the same participant.¹⁴ This double randomization generates clean variation for estimating substitution patterns (Berry and Haile, 2024).¹⁵

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 a detailed discussion).

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.¹⁶ Following the pre-registered sample selection criteria, we excluded 14% participants who failed comprehension questions, did not complete the survey, or spent less than one minute on the study, leaving a final sample of 9,265 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.

¹⁴Prices were drawn from a discrete uniform distribution ranging from \$3 to \$7.

¹⁵Although all participants technically see the same set of ten books in the first choice task, they rarely choose books that are displayed near the bottom of the list. Therefore, by randomizing the order in which books are displayed to participants, we vary their effective choice sets.

¹⁶The 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.

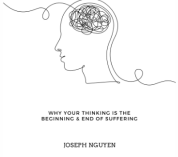
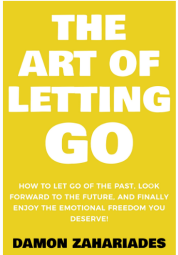
Book	Info	Reviews
	<p>Don't Believe Everything You Think by Joseph Nguyen Price: \$3 Genre: Self-Help Year: 2022 Pages: 126</p> <p>Description Learn how to overcome anxiety, self-doubt & self-sabotage without needing to rely on motivation or willpower. In this book, you'll discover the root cause of all psychological and emotional sufferin... read more</p>	<p>Read customer reviews: Review 1 Review 2 Review 3 Review 4 Review 5</p> <p>SELECT THIS BOOK</p>
	<p>The Art of Letting Go by Damon Zahariades Price: \$4 Genre: Self-Help Year: 2022 Pages: 196</p> <p>Description Finally Let Go of Your Negative Thoughts and Enjoy the Emotional Freedom You Deserve! Are you struggling with anger, regrets, and resentment? Do you feel emotionally exhausted, stressed, and discoura... read more</p>	<p>Read customer reviews: Review 1 Review 2 Review 3 Review 4 Review 5</p> <p>SELECT THIS BOOK</p>

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

3.2 Model Specifications and Estimation Results

We compare our approach against two benchmark models: the plain logit model and a standard logit model with random coefficients model on 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, 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.¹⁷ This variation gives us an additional exogenous shifter of choices, which helps us identify substitution patterns in a manner similar to how price variation does.

We estimate three demand models:

¹⁷We 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.

Model	Random Coefficients	AIC	ΔAIC
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

Table 1: **Comparison of models in terms of in-sample AIC on first choices.** The second column shows variables that have random coefficients in the selected specification. The last column shows the AIC reduction relative to plain logit.

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 .¹⁸ We perform model selection as detailed in Algorithm 1 in Section 2.
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 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 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

To evaluate demand models, we let them confront a difficult counterfactual prediction problem—predicting second choices, i.e. which books participants switch to when their

¹⁸Since 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.

most preferred option becomes unavailable. A model can only predict second choices well if it has accurately captured the true substitution patterns. Predicting second choices is also directly relevant in antitrust where second-choice *diversion ratios* are often used to measure substitutability and the intensity of price competition (Conlon and Mortimer, 2021).

Operationally, we first estimate each model via Maximum Likelihood Estimation (MLE) using first choices only. We then use the estimated model to predict counterfactual second-choice diversion ratios, $\hat{s}_{j \rightarrow k}$, 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, $s_{j \rightarrow k}$, computing *RMSE* as:

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

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

Figure 2 summarizes the validation results, while Appendix 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.

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

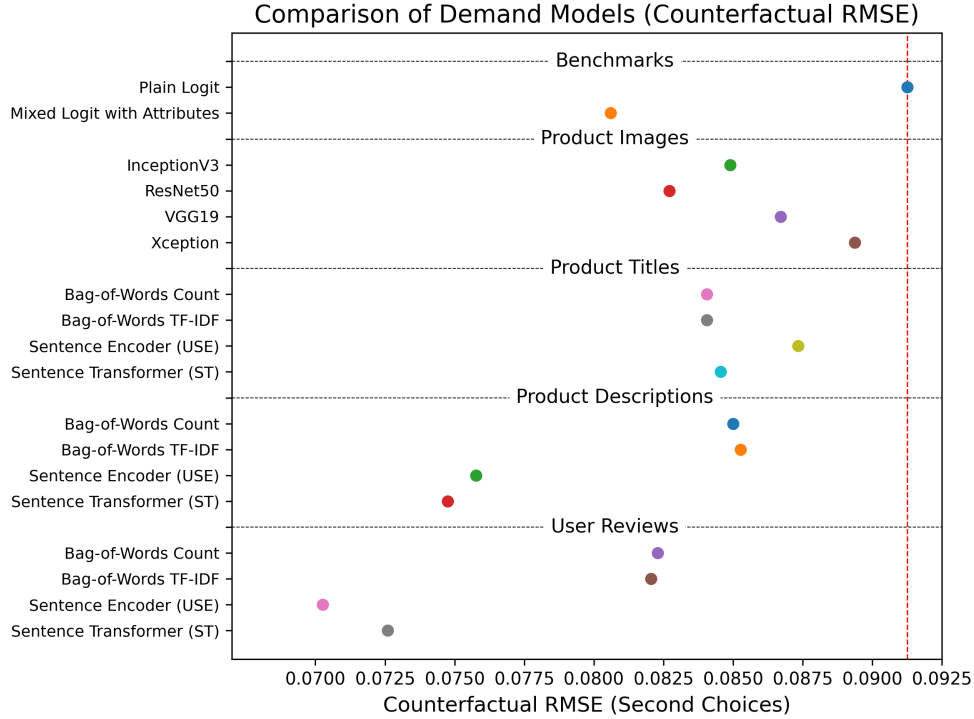


Figure 2: **Comparison of models in terms of counterfactual $RMSE$ on second choices.** The two benchmarks in the first panel are the plain logit without random coefficients and the best-fitting mixed logit with random coefficients on observed attributes. The remaining specifications correspond to mixed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

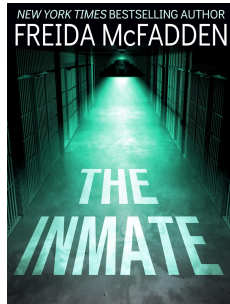
include symbolic elements such as skulls and swords that convey danger or peril, and feature natural objects like twisted vines and golden roses. Similarly, the self-help books have minimalistic layouts and consistent color schemes, such as black-and-white text on yellow backgrounds. Thus, image embeddings partly encode books' genres, which correlate with substitution patterns.

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

First, performance tends to improve when 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 describe them in 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, the USE model reduces $RMSE$ by 4.3% with book titles, 17% with descriptions, and 23% with

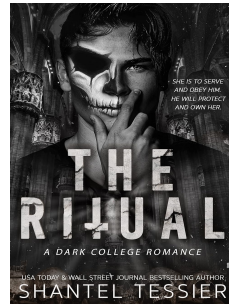
Mystery Books



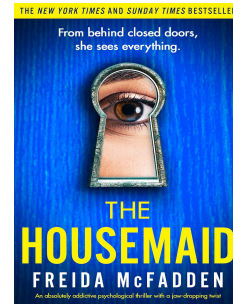
“The Inmate”



“Please Tell Me”



“The Ritual”

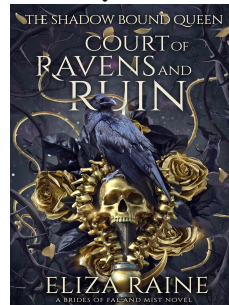


“The Housemaid”

Fantasy Books



“The Ashes & The Star Cursed King”

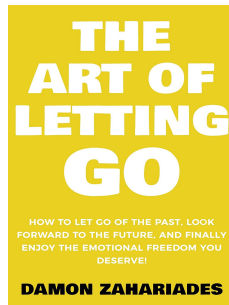


“Court of Ravens and Ruin”

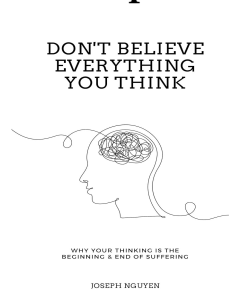


“The Serpent & The Wings of Night”

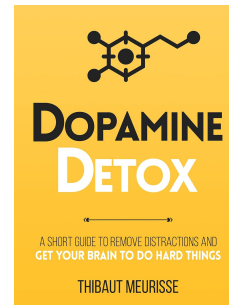
Self-Help Books



“The Art of Letting Go”



“Don't Believe Everything You Think”



“Dopamine Detox”

Figure 3: Ten books used in our experiment.

reviews relative to the plain logit. This makes intuitive sense. Although book titles are brief, they often contain subtle genre cues. For example, mystery titles frequently hint at secrets (*Please Tell Me*) or characters in confined situations (*Housemaid*, *Inmate*), whereas fantasy titles often include words that evoke a gothic atmosphere (*Serpent*, *Wings*, *Ravens*, *Ruin*). Descriptions provide even clearer genre signals, emphasizing unexpected twists such as hidden identities, past relationships, and betrayals. Reviews go even further: consumers summarize plots while also expressing opinions. When reviewing mystery novels, for instance, some readers praise cliffhangers and intriguing twists while others critique pacing issues like slow starts or rushed endings. Thus, it is unsurprising that reviews provide more information about substitution than descriptions, which, in turn, are more informative than titles.

Extension: Combining Data Types If text, images, and attributes provides 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. Neither extension improves *AIC* or *RMSE*. This result suggests information across data types is highly correlated, with text providing the strongest signal of substitution.

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-fitting *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.

Implied Substitution Patterns To illustrate how our approach improves predictions of substitution patterns, in Table 2 we compare predicted second-choice probabilities with

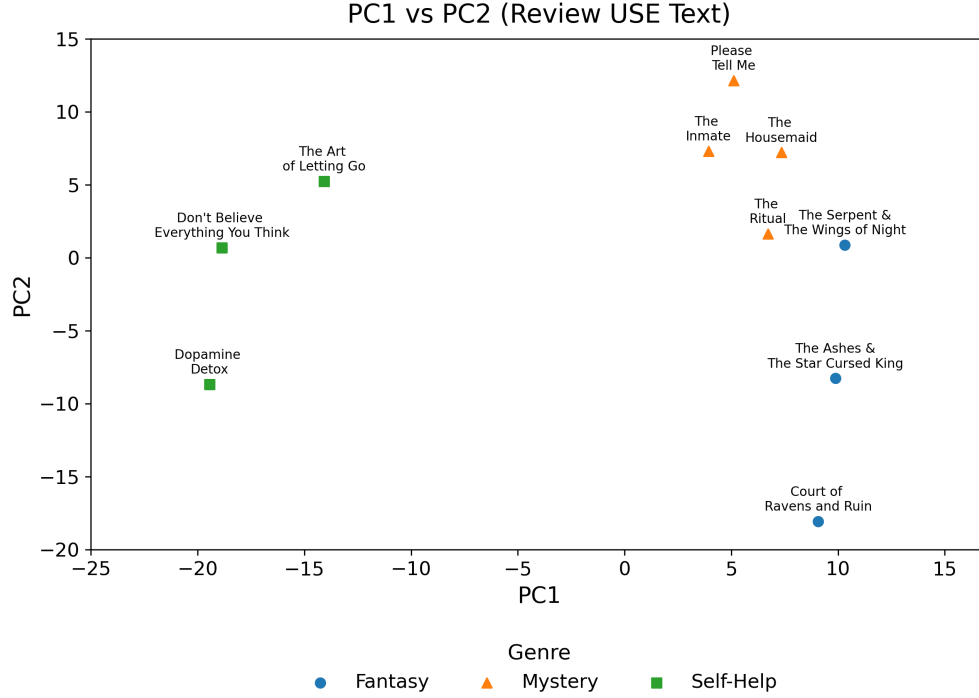


Figure 4: **Book locations in the space of selected principal components (Review USE model).**

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

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

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

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.

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. Further, *AIC* selects the specification that has the lowest second-choice *RMSE* across all considered specifications. 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.

3.4 Implications for Pricing: Merger Simulations

To illustrate how estimated substitution patterns can influence counterfactuals of interest, we conduct simulations of horizontal mergers—a natural application of our approach. Antitrust agencies routinely use demand models to assess whether a hypothetical merger would lead to a significant price increase ([Federal Trade Commission, 2022](#)). If the merging firms’ products are close substitutes, the merger creates strong upward pricing pressure, as the firm can “recapture” some consumers after raising prices. Therefore, predicting how a merged firm would set prices requires accurate estimates of substitution patterns.

For each pair of books, we compute their prices under two scenarios: (a) when the books are owned by separate publishers competing in a Bertrand-Nash equilibrium, and (b) when both books are owned by the same publisher setting prices by solving a joint first-order condition. In both cases, publishers take the prices of the other eight books as given, fixed at \$5, and have zero marginal costs.¹⁹

We recognize that limiting the choice set to ten books makes our analysis somewhat artificial, as major publishers typically manage vast assortments with hundreds of thousands of titles. Nevertheless, our simulations offer a natural way to evaluate how estimated substitution patterns translate into pricing decisions. This application is also

¹⁹We fix the prices of the other books at \$5, the average value in the experimental dataset. We do not optimize prices of other books because the model does not include an outside option. Consumer choices remain unchanged if all prices increase by the same amount, thus leading to multiple equilibria.

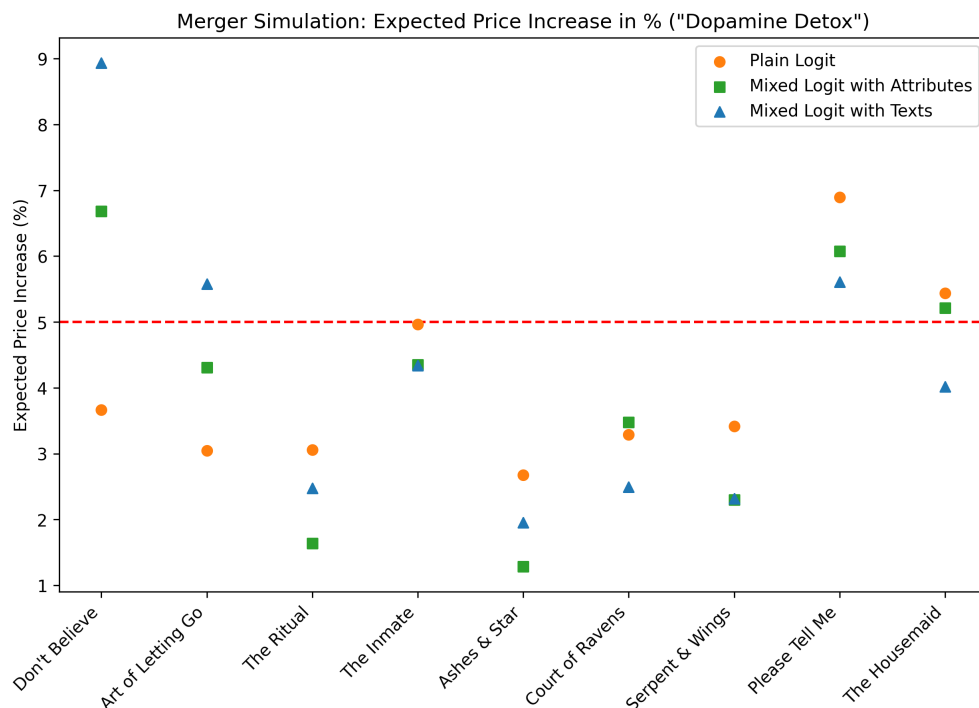


Figure 5: **Results of hypothetical merger simulations.** For each simulated merger of *Dopamine Detox* with one other book (horizontal axis), the figure shows the average price increase among the two merging books. The horizontal dashed line represents a hypothetical policy where the decision-maker challenges all mergers expected to raise prices by at least 5%. Appendix Table A2 reports the exact price increase estimates used to construct this graph.

policy-relevant given the high-profile mergers among major book publishers in the past few decades, which have drawn attention and regulatory scrutiny.²⁰

Figure 5 illustrates the predicted price increase resulting from the joint ownership of *Dopamine Detox* with each of the other books. We select *Dopamine Detox* as an example where our approach outperforms alternative models in capturing substitution patterns (see Table 2). For each simulated merger, Figure 5 reports the average price increase across the two merging books.

As discussed above, the plain logit model fails to capture strong within-genre substitution. Consequently, it incorrectly identifies as the closest substitutes for *Dopamine Detox* the three mystery books, rather than other self-help ones. This misclassification affects the predicted price changes in merger simulations. Specifically, the plain logit

²⁰In 2022, a federal judge blocked the proposed merger between *Penguin Random House* and *Simon & Schuster*, citing concerns that it would stifle competition (U.S. Department of Justice, 2022). This merger would have reduced the “Big Five” publishers to the “Big Four” and significantly increased the concentration of the publishing market. Prior to that, *Penguin Books* and *Random House* merged in 2013 to form *Penguin Random House*—the largest trade book publisher globally.

model overstates the price increase when *Dopamine Detox* is merged with the mystery book *Please Tell Me* or *The Housemaid*, while understating the price increase when merged with its true closest substitutes, *The Art of Letting Go* or *Don't Believe Everything You Think*. These discrepancies are substantial: for within-genre mergers, the plain logit predicts modest price increases of 3% and 3.7%, whereas our review-based model estimates them to be 5.6% and 8.9%—approximately 2 to 2.5 times higher (see Appendix Table A2).

The attribute-based mixed logit model produces predictions more aligned with the review-based model. In most cases, both models deviate from the plain logit in the same direction, yielding similar or even identical predicted price increases. However, notable discrepancies remain—for instance, in the case of the self-help books *The Art of Letting Go* or *Don't Believe Everything You Think*, the attribute-based model underestimates the price increase by approximately 25% compared to the review-based model.

These divergent predictions could lead decision-makers to different conclusions. As an example, consider an antitrust agency that applies a heuristic rule, challenging all mergers expected to increase prices by more than 5% (Bhattacharya et al., 2023). Compared to decisions made using the review-based model, which best captures substitution patterns, a decision-maker relying on the plain logit model would approve two mergers that should be challenged (*Don't Believe Everything You Think*, *The Art of Letting Go*) and challenge one that should be approved (*The Housemaid*). Similarly, a decision-maker using the attribute-based mixed logit model would approve a merger that should be challenged (*The Art of Letting Go*) and challenge one that should be approved (*The Housemaid*).

4 Application to Online Retail Data

Next, we apply the same 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 identify which text and image data best predict substitution patterns in various categories, thus offering practical guidance on what types of unstructured 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 Greninger 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

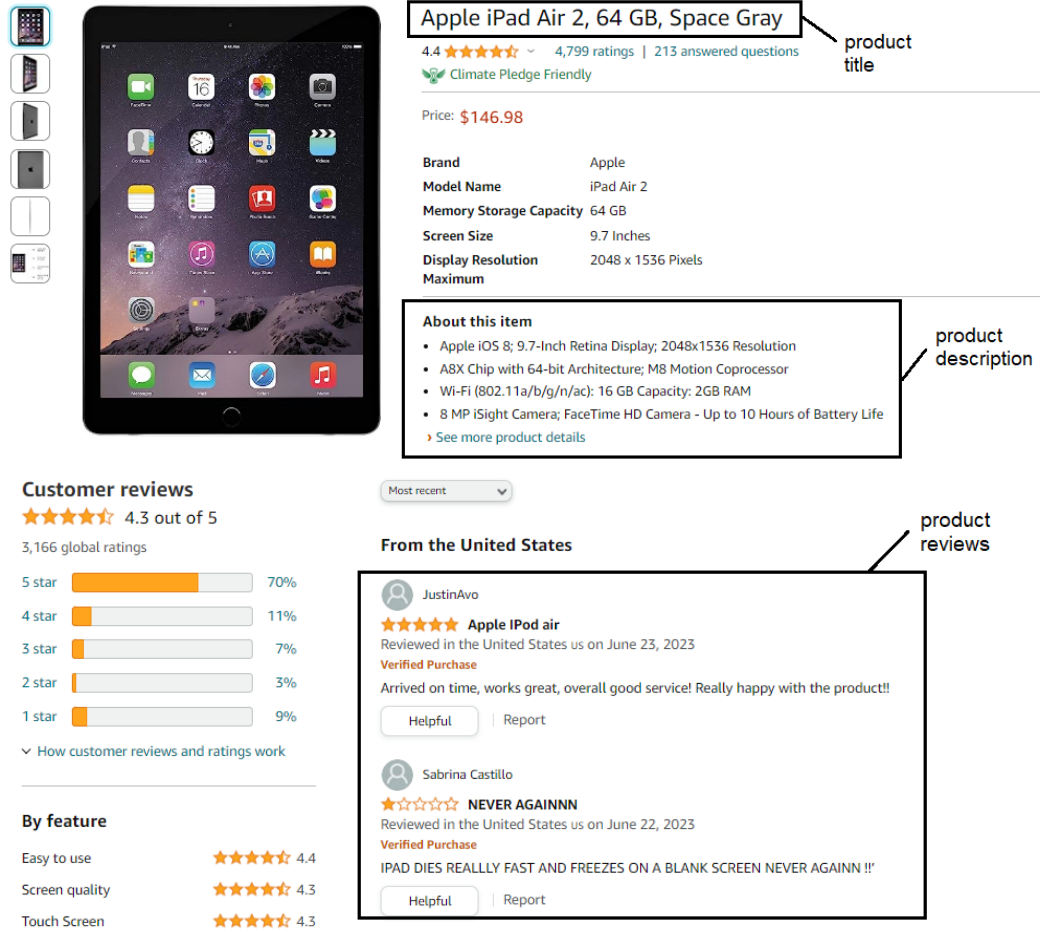


Figure 6: Example: image and text data collected from product detail pages.

price histories obtained from the third-party database Keepa.com.²¹ 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 6). 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 in response to unobserved demand shocks, we find that product images, titles, and descriptions very rarely change over time (see Appendix F).

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,

²¹We do not observe product rankings, so we do not include them in our demand models.

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 2,000 purchases. This criterion ensures that we observe enough purchases to estimate both product-fixed effects and substitution patterns.²² The average product price in selected categories is \$43.

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

4.2 Estimation Results

We apply our approach to each product category and compare it to the plain logit model. In the four electronics categories for which we collected attribute data, we also compare it to the attribute-based mixed logit model. Since we have a relatively large number of attributes in these categories, we reduce their dimensionality via PCA and apply Algorithm 1 for model selection.²³

Appendix Table A3 summarizes the estimation results, showing the selected model and data type for each category, while Appendix Figure A5 plots the distribution of *AIC* improvements relative to the plain logit. 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 5.0.²⁴ For the average category, our method reduces *AIC* by 23.3, with improvements reaching as high as 111.5 in some categories.²⁵ These results suggest that our approach consistently captures signals of substitution from text and images across a wide range of categories.

²²We omit products purchased fewer than 10 times and those with missing price data, which means some categories in our sample include fewer than 15 products.

²³We normalize each attribute to have mean zero and variance one before applying PCA.

²⁴When 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).

²⁵In 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.

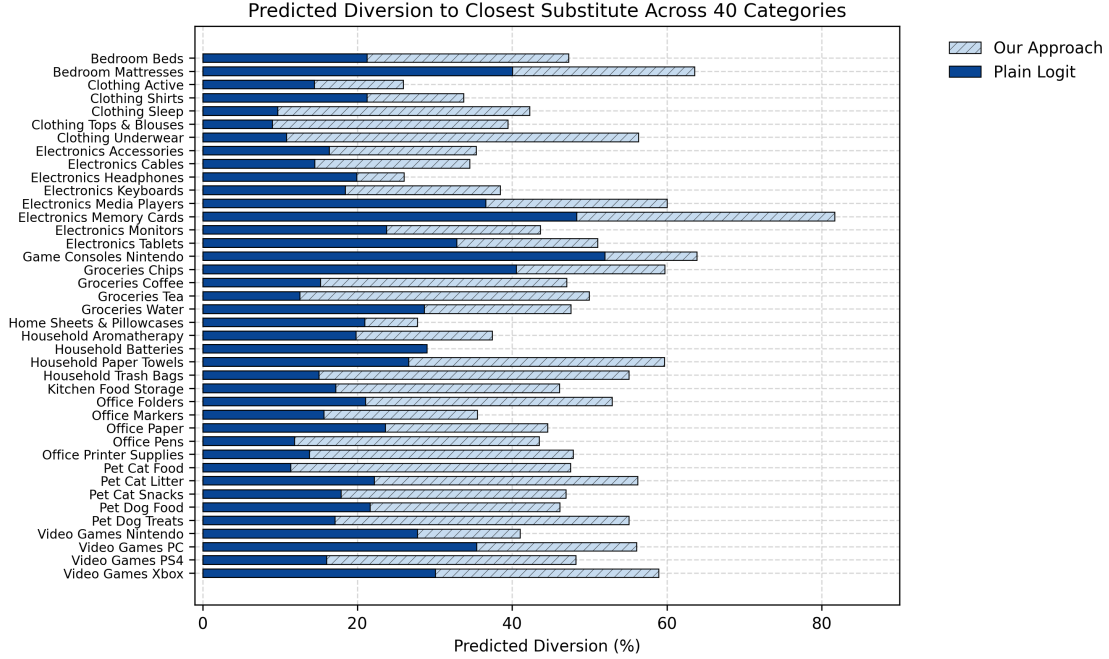


Figure 7: **Estimated diversion ratios to closest substitutes.** This figure plots the estimated diversion ratios to the closest substitutes, $\max_k \hat{s}_{j \rightarrow k}$, averaged across products j in each category.

Appendix Table A4 shows results from the electronics categories with attribute data. In all four categories, unstructured data meaningfully improves model fit relative to observed attributes, 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.

Next, we examine the estimated substitution patterns. While we do not observe the “ground truth” diversion ratios as we did in the experiment, we can still assess how estimated diversion ratios deviate from those in the plain logit model. A well-known limitation of plain logit is that diversion ratios depend only on market shares and not on attribute similarities, so this model often fails to identify close substitutes, producing overly flat diversions (Conlon et al., 2023). If our approach recovers substitution better, it should produce more variable diversion ratios.

Figure 7 plots the estimated diversion ratios to the closest substitutes, $\max_k \hat{s}_{j \rightarrow k}$, averaged across products j . Consistent with the intuition above, the plain logit yields relatively small diversion ratios, about 22% on average, while our approach increases this average substantially to 47%. In some categories, our approach estimates diversion ratios as high as 60-80%, which the plain logit does not produce. These results confirm that

our method generates significantly more variable diversion ratios, suggesting it better identifies which products are close substitutes.

Apart from showing greater variability, the diversion ratios predicted by our method are also more intuitive. For example, Appendix Tables A5 and A6 report predicted diversions for tablets. Because the plain logit relies on product fixed effects, it predicts that if any tablet is removed, consumers substitute to the most popular alternatives (e.g., Fire 7 or Fire HD10), producing unrealistically flat diversions. By contrast, the *Description ST* model chosen by our model selection algorithm yields more intuitive diversions: the two kids’ tablets are estimated to be close substitutes (Fire Kids and Fire HD8 Kids), the two iPads are close substitutes (iPad 9.7 and iPad 10.2), and so on.

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 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, for each data type d (images, titles, descriptions, or reviews) we compute the Akaike weight w_d as

$$w_d = \frac{\sum_{\{r:d(r)=d\}} \exp(-\Delta_r/2)}{\sum_i \exp(-\Delta_i/2)} \quad (3)$$

where i indexes estimated specifications, and $\Delta_i = AIC_i - AIC_{min}$ where AIC_{min} is the lowest *AIC* across *all* estimated specifications in that product category. The numerator in (3) sums over estimated specifications using data type d , while the denominator sums over all specifications estimated during model selection. 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.²⁶

Figure 8 shows the estimated Akaike weights by category and data type. We find considerable variation across categories in terms of which types of data are most important for predicting substitution patterns.

Importantly, many of these results are difficult to predict in advance. For example, while we might expect visual features to be most relevant in clothing categories, where consumers care about visual design, only two of the five clothing categories (“Activewear”

²⁶More 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, p.272).

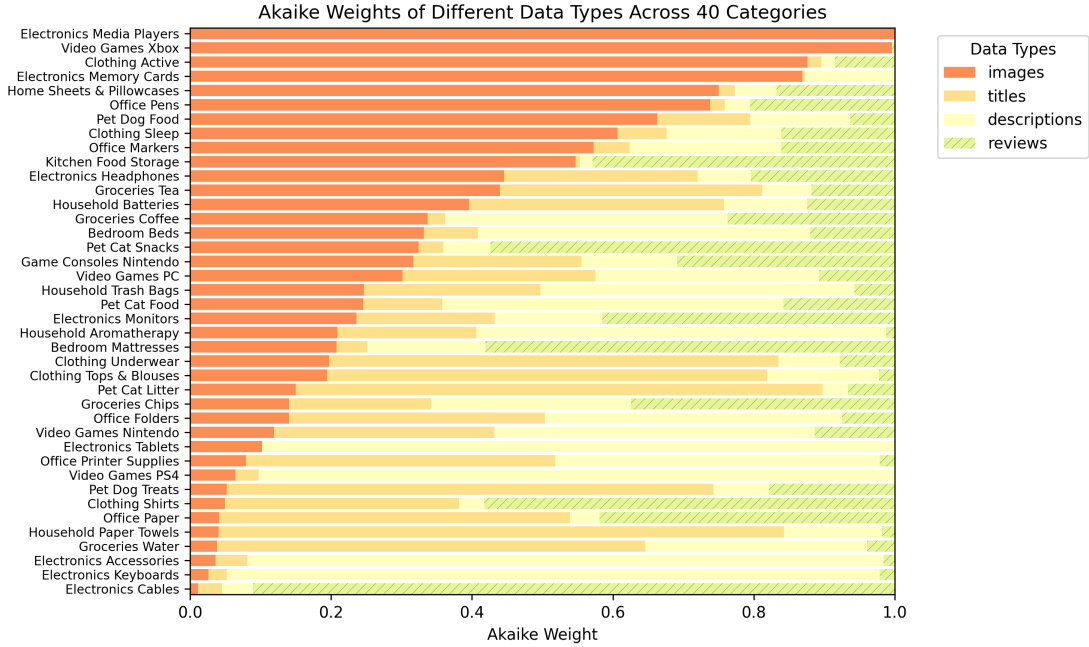


Figure 8: **Akaike Weights of different types of unstructured data across 40 product categories.** These weights reflect the relative importance of data types for predicting substitution patterns in the data. We compute Akaike weights using the formula in (3).

and “Sleep”) assign more weight to images than to text. In contrast, for “Tops & Blouses” and “Underwear,” product titles are most predictive of substitution, whereas for “Shirts,” reviews perform best. Similarly, while we may expect descriptions and reviews of video games to be as informative as those for books, the data strongly suggests that in the category of video games for Xbox, images contain substantially more information about substitution patterns than text.

These results highlight that researchers cannot reliably predict in advance which data types will best capture substitution. Therefore, in practice, it is important to collect different data types and use model selection to identify the one that is most predictive of 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 identified the specification that delivered 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 did 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. This can be done by checking whether adding some subset of observed attributes to the selected model with unstructured data further reduces *AIC*. When many attributes are available, researchers can reduce their dimensionality using PCA, as in Section 4.2.

5.2 Counterfactual Analysis and its Boundaries

We interpret embeddings as capturing time-invariant product aspects (e.g., attributes, functional benefits, and visual designs). This assumption is reasonable given that in our application, product images, titles, and descriptions rarely change over time (see Appendix F). Thus, our method is well-suited for a wide range of counterfactuals that require estimating consumer responses to price changes conditional on the other product attributes, including optimal pricing, simulating merger effects on prices, evaluating corrective taxes, or estimating markups (Berry and Haile, 2024). In such contexts, researchers can evaluate counterfactual market shares and prices just as in standard demand models, assuming that embeddings—and the resulting demand functions—do not change. Moreover, as usual, researchers can quantify the welfare effects of price changes by calculating the area under the demand curve (Small and Rosen, 1981). The fact that principal components have no direct interpretation does not preclude such analysis.

Our approach can also be applied to study consumer responses to changes in product availability, for example, to quantify consumer surplus from new products or study how firms optimize assortments. In retrospective analyses, where the introduction of new

products is observed in the data as in [Petrin \(2002\)](#), researchers can use our approach “as is” knowing that it performs well at predicting consumer responses to product removals. By contrast, evaluating a hypothetical new product that is not yet in the data poses an additional challenge: researchers need to obtain images and textual descriptions of this product from pre-launch information.

In some applications, embeddings may change in counterfactual scenarios. For example, if customer reviews mention value relative to price (e.g., “*Great quality for the price!*” or “*Good product, but overpriced.*”), then the reviews and the resulting text embeddings will shift when prices change in counterfactuals. Researchers can diagnose this issue by checking how common such reviews are in their data and, if necessary, model how price changes affect embeddings.

Another example is merger simulations: if firms redesign or reposition products after merging, this may alter embeddings and hence substitution patterns. If this is a first-order concern, researchers may want to specify a supply-side model in which firms choose both prices and non-price characteristics ([Fan, 2013](#)).

That said, our method does allow for correlation between embeddings and variables that change in counterfactuals. For instance, embeddings may be correlated with prices if they capture unobserved product quality. This does not rule out counterfactuals where prices change (e.g., due to tariffs), since such changes do not alter the product’s intrinsic quality.

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 robustly 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.²⁷

Second, future research should provide a formal treatment of how to perform inference in a way that accounts for the uncertainty in the model selection step. One possibility would be to use sample-splitting as in [Wasserman and Roeder \(2009\)](#) and [Taylor and Tibshirani \(2015\)](#), randomly selecting a subset of the data for model selection and using the other for estimation and inference based on the selected model.

Third, other demand models may leverage text and image data more effectively than the random coefficients logit model. In an earlier version of this paper, we tested the pairwise combinatorial logit model of [Koppelman and Wen \(2000\)](#), which allows each product pair to have its own utility correlation based on distance in the embedding space. The model, however, performed significantly worse than our mixed logit, not only when using unstructured data but also with observed attributes, where we let utility correlations depend on the distance between products in the attribute space. This highlights how the functional form assumptions of the demand model can significantly impact counterfactual performance and suggests that combining unstructured data with more flexible demand models, such as that in [Compiani \(2022\)](#), may be a fruitful direction for future research. We note that this point applies more generally to any discrete choice model and is not specific to our method.

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 Amazon application, we rely on product fixed effects to capture unobserved attributes (e.g., quality) that may correlate with prices. More generally, researchers may worry that prices 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](#)).²⁸

²⁷Replication codes and experimental data are available in our public repository: github.com/ilyamorozov/DeepLogitReplication. Separately, the Python package for implementing our method is available on PyPI and at: github.com/deep-logit-demand/deeplogit.

²⁸For 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.

6 Conclusion

In this paper, we demonstrate 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.

References

- AKAIKE, H. (1998): “Information theory and an extension of the maximum likelihood principle,” in *Selected papers of hirotugu akaike*, Springer, 199–213.
- 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,” .
- BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): “Automobile Prices in Market Equilibrium,” *Econometrica*, 63, 841–890.
- (1999): “Voluntary export restraints on automobiles: Evaluating a trade policy,” *American Economic Review*, 89, 400–431.
- (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.
- BHATTACHARYA, V., G. ILLANES, AND D. STILLERMAN (2023): “Merger Effects and Antitrust Enforcement: Evidence from US Consumer Packaged Goods,” Tech. rep., National Bureau of Economic Research.
- 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, N. HUA, N. LIMTIACO, R. S. JOHN, N. CONSTANT, M. GUAJARDO-CESPEDES, S. YUAN, C. TAR, Y.-H. SUNG, B. STROPE, AND R. KURZWEIL (2018): “Universal Sentence Encoder,” .
- CHOLLET, F. (2017): “Xception: Deep Learning With Depthwise Separable Convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258.

- 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 (2022): “Estimating preferences and substitution patterns from second choice data alone,” Tech. rep., working paper.
- (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.
- FAN, Y. (2013): “Ownership consolidation and product characteristics: A study of the US daily newspaper market,” *American Economic Review*, 103, 1598–1628.
- FEDERAL TRADE COMMISSION (2022): “Demand System Estimation and Its Application to Horizontal Merger Analysis,” Tech. rep., Federal Trade Commission, accessed: February 12, 2025.
- GOLDBERG, P. K. (1995): “Product differentiation and oligopoly in international markets: The case of the US automobile industry,” *Econometrica: Journal of the Econometric Society*, 891–951.
- GOODFELLOW, I., Y. BENGIO, A. COURVILLE, AND Y. BENGIO (2016): *Deep learning*, vol. 1, MIT press Cambridge.
- 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 Cambridge, Mass., USA.

- 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.
- 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, D. CER, M. SHANBHOGUE, I. NAIM, G. H. ÁBREGO, Z. LI, K. CHEN, H. S. VERA, ET AL. (2025): “Gemini embedding: Generalizable embeddings from gemini,” *arXiv preprint arXiv:2503.07891*.
- LEE, K. (2024): “Generative Brand Choice,” .
- MAGNOLFI, L., J. MCCLURE, AND A. SORENSEN (2022): “Triplet Embeddings for Demand Estimation,” Working Paper.
- MCCLURE, J. (2025): “Using Default Recommendations in Demand Estimation,” Working paper, Purdue University, Mitch Daniels School of Business.
- 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*.
- NETZER, O., R. FELDMAN, J. GOLDENBERG, AND M. FRESKO (2012): “Mine Your Own Business: Market-Structure Surveillance Through Text Mining,” *Marketing Science*, 31, 521–543.
- NEVO, A. (2000): “Mergers with differentiated products: The case of the ready-to-eat cereal industry,” *The RAND Journal of Economics*, 395–421.
- (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, J. KRAUSE, S. SATHEESH, S. MA, Z. HUANG, A. KARPATY, A. KHOSLA, M. BERNSTEIN, A. C. BERG, AND L. FEI-FEI (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.
- SHAPIRO, C. (1995): “Mergers with differentiated products,” *Antitrust*, 10, 23.
- 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*, forthcoming.
- SMALL, K. A. AND H. S. ROSEN (1981): “Applied welfare economics with discrete choice models,” *Econometrica: Journal of the Econometric Society*, 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.
- TAYLOR, J. AND R. J. TIBSHIRANI (2015): “Statistical learning and selective inference,” *Proceedings of the National Academy of Sciences*, 112, 7629–7634.
- U.S. DEPARTMENT OF JUSTICE (2022): “Justice Department Obtains Permanent Injunction Blocking Penguin Random House’s Proposed Acquisition of Simon & Schuster,” Accessed: February 9, 2025.
- VILCASSIM, N. J. AND P. K. CHINTAGUNTA (1995): “Investigating retailer product category pricing from household scanner panel data,” *Journal of Retailing*, 71, 103–128.
- WASSERMAN, L. AND K. ROEDER (2009): “High dimensional variable selection,” *Annals of statistics*, 37, 2178.
- ZHANG, Y., M. LI, D. LONG, X. ZHANG, H. LIN, B. YANG, P. XIE, A. YANG, D. LIU, J. LIN, ET AL. (2025): “Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models,” *arXiv preprint arXiv:2506.05176*.

Online Appendix

A Text Processing Steps

Before applying our text models, listed in Section 2, we pre-process our text data as follows. Working with product titles is straightforward because each title is a short text that typically includes only 10-15 words. When assessing similarity based on product descriptions, we merge the text from all bullet points, and apply our models to the merged text. For customer reviews, we transform the text of each review into a separate vector of word occurrences or an embedding, and we average these vectors or embeddings across all reviews of a given product.

For both bag-of-words approaches, we further pre-process text data by removing stopwords and lemmatizing words. We remove stopwords using the standard dictionary of common English words in the NLTK package. We lemmatize words using the WordNet Lemmatizer from the same package, NLTK. Then, we convert each pre-processed text into a vector of word occurrences (weighted word occurrences for TF-IDF). The other two models, USE and ST, have a built-in text pre-processing step. We therefore apply these models directly to the unprocessed text data.

B Book Selection Procedure

Which books we show to participants can significantly affect the performance of different demand models. Our general strategy was to select a set of books that does not bias our model comparisons in favor of any particular specification. To this end, we aimed to maximize variation in choice-relevant structured attributes as well as in text and image embeddings.

We chose books from three genres—Mystery, Fantasy, and Self-Help—and identified 20 books of each genre from Amazon’s bestseller lists. We included books from three different genres in order to generate meaningful variation in this structured attribute, which we anticipated would be predictive of substitution patterns.

To avoid arbitrary book selection within each genre, we implemented the following algorithm. We considered all possible combinations of ten books such that the three genres are roughly equally represented. Among these, we choose the combination of books that maximized the variance of text and image embeddings.²⁹ The final set, shown

²⁹Operationally, we computed the variance of image embeddings (using the VGG19 model) and the variance of text embeddings (using the USE model), then averaged the two. We performed a brute-force search

in Figure 3, includes four mystery books, three fantasy books, and three self-help books.

C Choice Survey: Sanity Checks

Because choices in our experiment were hypothetical and not incentivized, we want to verify that participants made meaningful choices. Several summary statistics suggest that participants took the choice tasks seriously. First, only 50 of the initially recruited participants (less than 1%) completed the entire study in less than one minute and thus had to be dropped from our sample. The remaining participants spent, on average, 7 minutes on the survey overall and 1.3 minutes on the choice tasks, indicating they took their time to make careful selections and did not mindlessly click through the survey (Figure A1).

Second, in the choice tasks, participants were disproportionately more likely to select books of the genre they reported to be their favorite in a questionnaire before the choice tasks (Figure A2), suggesting they considered the book attributes when making their choices.

Finally, participants' choices were consistent across the two choice tasks (Figure A3). For example, over 60% of participants who selected a mystery book in the first task chose another mystery book in the second choice task. This observation suggests participants considered their genre preferences when making both choices.

over all possible sets of ten books and selected the set that maximized this average variance.

D Computing Second-Choice Diversion Ratios

To compute $RMSE$ in (2), we need to calculate predicted second-choice diversion ratios $\hat{s}_{j \rightarrow k}$ and their data counterparts $s_{j \rightarrow k}$. Recall that both $s_{j \rightarrow k}$ and $\hat{s}_{j \rightarrow k}$ reflect the probability that the consumer chooses book k in the second choice task conditional on having chosen book j in the first choice task. Using the analogy principle, we compute empirical diversions $s_{j \rightarrow k}$ using a simple frequency estimator:

$$s_{j \rightarrow k} = \frac{\sum_{i=1}^N \mathbf{1}\{y_i^{(2)} = k, y_i^{(1)} = j\}}{\sum_{i=1}^N \mathbf{1}\{y_i^{(1)} = j\}} \quad \text{for } j \neq k, \quad (4)$$

where $y_i^{(1)}$ and $y_i^{(2)}$ are participant i 's first and second choices. By contrast, to compute diversions $\hat{s}_{j \rightarrow k}$ predicted by a given demand model, we use the identity

$$Pr(y_i^{(2)} = k | y_i^{(1)} = j; w_i) = \frac{s_k^{(j)}(w_i) - s_k(w_i)}{s_j(w_i)}, \quad (5)$$

where $s_j(w_i)$ and $s_k(w_i)$ are the unconditional choice probabilities of books j and k , $s_k^{(j)}(w_i)$ is the probability of choosing book k after book j is removed from the choice set, and w_i is a vector of prices ($\text{price}_{i1}, \dots, \text{price}_{iJ}$) and rankings ($\text{rank}_{i1}, \dots, \text{rank}_{iJ}$) that participant i encounters in the experiment. We average diversions in (5) across all participants to compute $\hat{s}_{j \rightarrow k}$:

$$\hat{s}_{j \rightarrow k} = \frac{1}{N} \sum_{i=1}^N \frac{s_k^{(j)}(w_i) - s_k(w_i)}{s_j(w_i)}. \quad (6)$$

E Attributes Collected for Electronics Products

For each of the four electronics categories described in Section 4.1, we collect standard attributes from Amazon product detail pages. Specifically, we include all specifications listed in the product descriptions and, when available, those in the ‘‘Technical Details’’ section. Below is the complete list of all attributes for each category:

1. **Tablets:** brand, model, memory, RAM, processor speed, number of cores, screen size, maximum resolution, charging time, battery life, front camera, back camera, front camera megapixels, back camera megapixels, warranty, whether the product comes with a case, and whether it includes an Amazon Kids subscription.

2. **Monitors:** brand, screen size, maximum resolution, refresh rate, blue light filter, frameless design, tilt adjustment, height adjustment, flicker-free display, built-in speaker, wall-mountable option, curved screen, and adaptive sync.
3. **Memory Cards:** brand, pack size, micro card, flash memory type, capacity, read speed, write speed, speed class, UHS speed class, device compatibility (smartphone, computer, camera, laptop, tablet), X-ray proof, temperature proof, waterproof, shock-proof, magnetic proof, and whether an adapter is included.
4. **Headphones:** brand, color, connectivity, number of eartip sets, battery life on a single charge, battery life with charging case, deep bass feature, tangle-free design, waterproof, sweat-proof, IPX rating, whether the product comes with a case, a microphone, and a noise reduction option.

F Changes in Text and Images Over Time

As discussed in Section 2, we treat text and image embeddings, as well as the principal components extracted from them, as being fixed over time for a given product. To validate this assumption, we examine whether text and images change over time.

Since we lack data on these changes for 2019-2020, we construct a separate sample by repeatedly collecting unstructured data from Amazon’s product detail pages daily from January 23 to March 4, 2025. To keep data collection manageable, we do not gather customer reviews and select a subset of 11 out of 40 categories, ensuring they cover all of Amazon’s departments (e.g., “Clothing,” “Food,” “Electronics”) represented in the full dataset.³⁰ The selected categories are: “Shirts,” “Coffee,” “Aromatherapy,” “Mattresses,” “Markers,” “Pet Litters,” “Nintendo Games,” “Tablets,” “Memory Cards,” “Monitors,” and “Earbuds.” In each category, we collect data for the products used in our estimation in Section 4 that were not discontinued, totaling 136 products.

We find that product images do not change over time. Titles change for only six products (4%), mostly by adding or removing specific attributes or functional benefits. Similarly, descriptions change for just 21 products (15%). Most changes do not affect which product features are revealed, but they alter which ones are immediately visible versus being revealed only after clicking on “See more product details.” Thus, we conclude that

³⁰Recall that we average over embeddings extracted from the 100 most recent reviews. Even though customer reviews accumulate over time as consumers continue to write them, the average embeddings extracted from these reviews may remain approximately constant if consumers consistently discuss the same attributes. While we do not have review data to verify this claim, future research should examine whether review embeddings are stable in online markets.

changes in text and images are not a significant concern for the products in our empirical application.

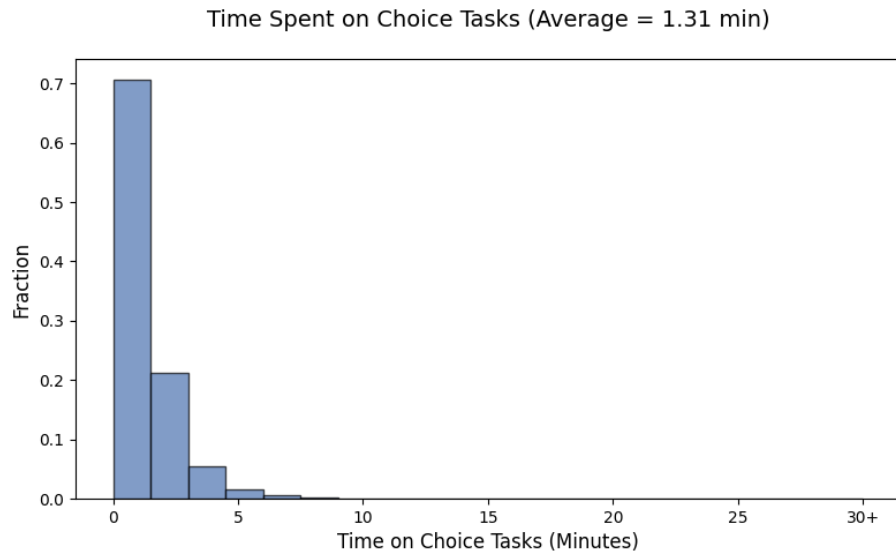
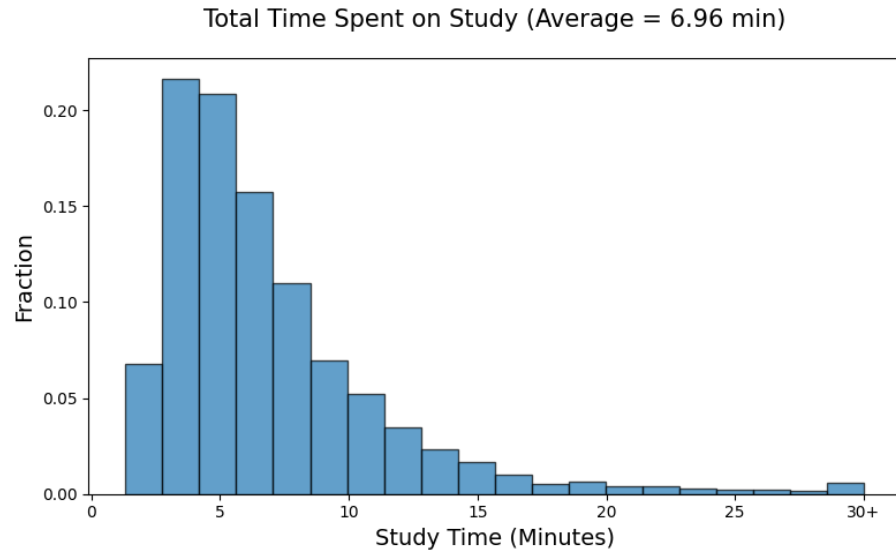


Figure A1: Time spent by participants on choice tasks in the experiment.

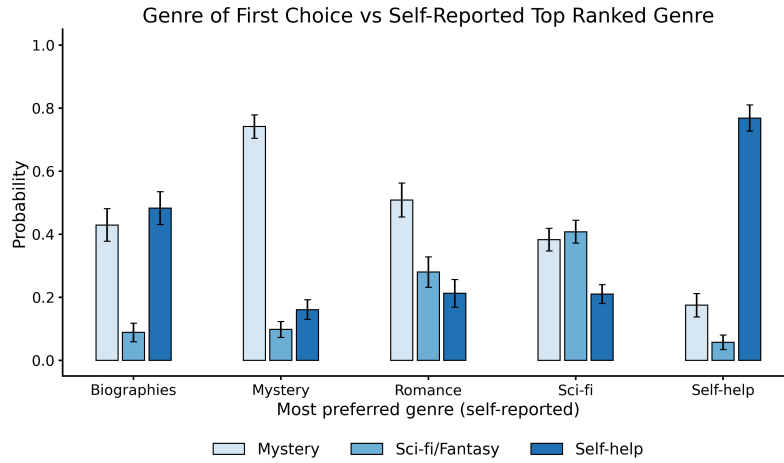


Figure A2: Selected genres and self-reported genre preferences in the experiment.

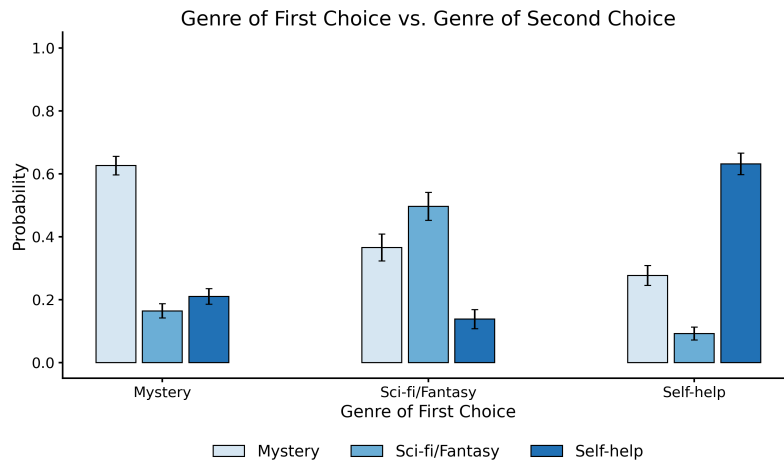


Figure A3: Genres of participants' first and second choices in the experiment.

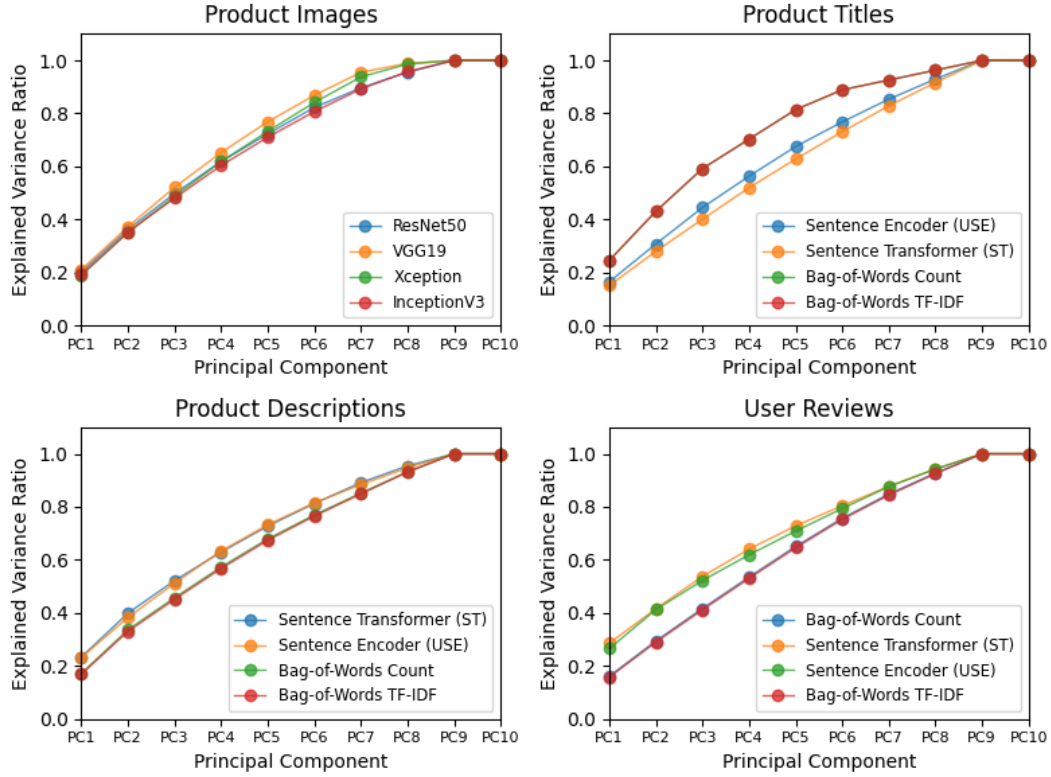


Figure A4: Share of variance of embeddings explained by principal components.

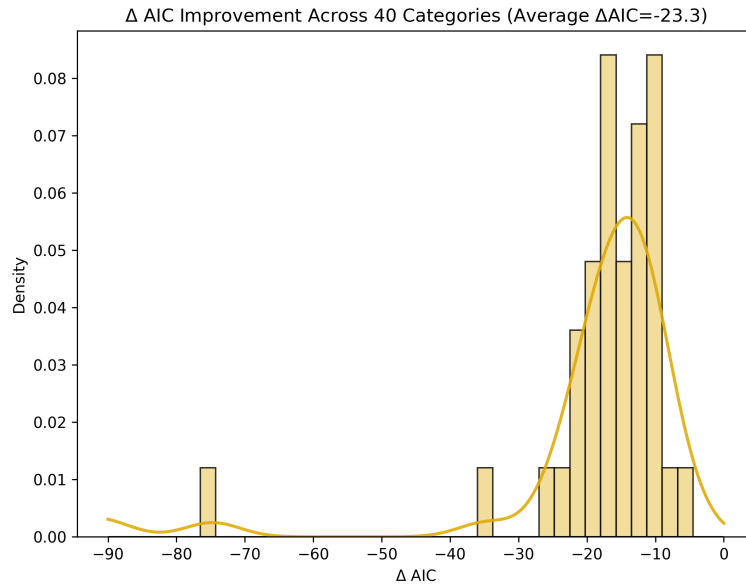


Figure A5: Fit improvements relative to plain logit across 40 product categories. Δ AIC in each category is the difference between the AIC of the selected specification with unstructured data and that of the plain logit model.

Model	First Choices (Data)		Second Choices (Counterf.)		
	$\log L$	AIC	RMSE	Rel. to Plain Logit Δ	%
Panel A. Benchmark Model					
Plain Logit	-20488.4	41006.7	0.091		
Panel B. Mixed Logit Models with Principal Components					
Images: InceptionV3	-20475.2	40990.5	0.085	-0.006	-7.0%
Images: ResNet50	-20479.9	40997.8	0.083	-0.009	-9.4%
Images: VGG19	-20481.2	41000.4	0.087	-0.005	-5.0%
Images: Xception	-20479.1	40996.3	0.089	-0.002	-2.1%
Product Titles: Bag-of-Words Count	-20478.7	40995.5	0.084	-0.007	-7.9%
Product Titles: Bag-of-Words TF-IDF	-20478.7	40995.5	0.084	-0.007	-7.9%
Product Titles: Sentence Encoder (USE)	-20481.2	40998.3	0.087	-0.004	-4.3%
Product Titles: Sentence Transformer (ST)	-20477.1	40992.3	0.085	-0.007	-7.3%
Descriptions: Bag-of-Words Count	-20475.5	40988.9	0.085	-0.006	-6.9%
Descriptions: Bag-of-Words TF-IDF	-20475.8	40989.6	0.085	-0.006	-6.5%
Descriptions: Sentence Encoder (USE)	-20474.2	40986.4	0.076	-0.015	-17.0%
Descriptions: Sentence Transformer (ST)	-20474.9	40987.8	0.075	-0.016	-18.1%
Reviews: Bag-of-Words Count	-20479.6	40997.1	0.082	-0.009	-9.8%
Reviews: Bag-of-Words TF-IDF	-20479.4	40996.9	0.082	-0.009	-10.1%
Reviews: Sentence Encoder (USE)	-20472.0	40981.9	0.070	-0.021	-23.0%
Reviews: Sentence Transformer (ST)	-20473.3	40984.6	0.073	-0.019	-20.4%
Panel C. Mixed Logit with Observed Attributes					
Price	-20485.9	41003.9	0.091	-0.000	-0.0%
Pages	-20484.9	41001.7	0.086	-0.005	-5.5%
Year	-20484.7	41001.4	0.091	0.000	0.1%
Genre	-20483.3	41000.7	0.084	-0.007	-7.7%
Price & Pages	-20478.9	40991.9	0.081	-0.010	-11.4%
Price & Year	-20484.1	41002.2	0.092	0.000	0.3%
Price & Genre	-20483.2	41002.5	0.086	-0.005	-5.2%
Pages & Year	-20478.3	40990.7	0.081	-0.011	-11.7%
Pages & Genre	-20481.6	40999.2	0.081	-0.010	-11.4%
Year & Genre	-20483.3	41002.7	0.084	-0.007	-7.7%
Price, Pages, & Year	-20478.3	40992.7	0.081	-0.011	-11.7%
Price, Pages, & Genre	-20478.9	40995.9	0.081	-0.010	-11.4%
Price, Year, & Genre	-20482.2	41002.4	0.084	-0.007	-7.4%
Pages, Year, & Genre	-20478.3	40994.7	0.081	-0.011	-11.7%
Price, Pages, Year, & Genre (All Attr.)	-20476.9	40993.9	0.078	-0.013	-14.1%

Table A1: **Model validation results.** The table shows in-sample fit on first-choice data and counterfactual performance on second-choice data for all specifications considered in Figure 2 (see Section 3.2 for detailed description of these models).

Book	Plain Logit	Mixed Logit with Attributes	Mixed Logit with Texts
Don't Believe	3.7%	6.7%	8.9%
Art of Letting Go	3.0%	4.3%	5.6%
The Ritual	3.1%	1.6%	2.5%
The Inmate	5.0%	4.4%	4.3%
Ashes & Star	2.7%	1.3%	2.0%
Court of Ravens	3.3%	3.5%	2.5%
Serpent & Wings	3.4%	2.3%	2.3%
Please Tell Me	6.9%	6.1%	5.6%
The Housemaid	5.4%	5.2%	4.0%

Table A2: **Merger Simulation Results (Dopamine Detox)**. The table shows predicted relative price increases resulting from a simulated merger between Dopamine Detox and each other book respectively. For each simulated merger between *Dopamine Detox* and another book (first column), the table reports the expected average price increase for the two merging books across three estimated demand models.

Category	Data Type	Model Type	Δ AIC	Δ Diversion to Closest Substitute
1. Clothing Active	Images	VGG16	-25.1	11.5%
2. Clothing Shirts	Reviews	USE	-11.9	12.5%
3. Clothing Underwear	Titles	TFIDF	-16.3	45.5%
4. Clothing Sleep	Images	Inceptionv3	-13.1	32.5%
5. Clothing Tops & Blouses	Titles	USE	-12.3	30.4%
6. Electronics Cables	Reviews	TFIDF	-16.7	20.1%
7. Electronics Accessories	Descriptions	COUNT	-19.6	19.0%
8. Electronics Keyboards	Descriptions	ST	-19.5	20.0%
9. Electronics Memory Cards	Images	VGG16	-96.8	33.4%
10. Electronics Tablets	Descriptions	ST	-111.5	18.2%
11. Electronics Monitors	Images	Resnet50	-22.4	19.9%
12. Electronics Headphones	Images	VGG19	-10.7	6.1%
13. Electronics Media Players	Images	VGG19	-90.2	23.4%
14. Groceries Water	Titles	USE	-19.1	19.0%
15. Groceries Coffee	Descriptions	TFIDF	-9.5	31.8%
16. Groceries Tea	Images	Xception	-17.1	37.4%
17. Groceries Chips	Descriptions	ST	-5.6	19.2%
18. Household Aromatherapy	Descriptions	COUNT	-11.2	17.6%
19. Household Batteries	Images	VGG16	-10.7	-0.6%
20. Household Trash Bags	Descriptions	TFIDF	-16.2	40.1%
21. Household Paper Towels	Titles	TFIDF	-16.9	33.1%
22. Home Sheets & Pillowcases	Images	Resnet50	-21.3	6.8%
23. Bedroom Beds	Descriptions	USE	-11.2	26.0%
24. Bedroom Mattresses	Reviews	ST	-12.9	23.5%
25. Kitchen Food Storage	Images	VGG19	-20.5	28.9%
26. Office Folders	Descriptions	ST	-17.4	31.9%
27. Office Paper	Reviews	ST	-13.5	21.0%
28. Office Markers	Images	Inceptionv3	-13.6	19.8%
29. Office Pens	Images	VGG19	-13.5	31.6%
30. Office Printer Supplies	Titles	USE	-15.1	34.1%
31. Pet Cat Food	Descriptions	COUNT	-11.1	36.2%
32. Pet Cat Litter	Titles	ST	-12.1	34.1%
33. Pet Cat Snacks	Reviews	COUNT	-16.9	29.1%
34. Pet Dog Food	Images	Inceptionv3	-11.7	24.5%
35. Pet Dog Treats	Titles	USE	-18.2	38.0%
36. Game Consoles Nintendo	Images	Resnet50	-6.8	11.9%
37. Video Games Nintendo	Descriptions	ST	-23.5	13.3%
38. Video Games PC	Titles	COUNT	-9.5	20.7%
39. Video Games PS4	Descriptions	ST	-34.7	32.2%
40. Video Games Xbox	Images	Xception	-74.7	28.9%
Averaged			-23.3	24.6%

Table A3: **Estimation results across 40 categories in Comscore data.** For each category, the table shows the selected model and data type yielding the lowest AIC and the AIC improvement relative to plain logit. The last column shows the increase in the predicted diversion ratios to the closest substitutes, $\max_k \hat{s}_{j \rightarrow k}$, averaged across products j , relative to plain logit (in percentage points).

	<i>AIC</i>	ΔAIC 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

Table A4: *AIC* Comparison of Our Approach and Mixed Logit with Attributes

	Fire 7	Fire HD 8	Fire HD 10	Fire 7 Kids	iPad 10.2	Fire HD 8 Kids	Dragon Touch	iPad 9.7
Fire 7	0.000	0.330	0.381	0.301	0.312	0.283	0.279	0.286
Fire HD 8	0.225	0.000	0.229	0.175	0.178	0.165	0.162	0.166
Fire HD 10	0.386	0.340	0.000	0.314	0.320	0.297	0.291	0.295
Fire 7 Kids	0.123	0.105	0.128	0.000	0.100	0.091	0.089	0.091
iPad 10.2	0.153	0.129	0.148	0.122	0.000	0.113	0.112	0.114
Fire HD 8 Kids	0.043	0.038	0.049	0.035	0.035	0.000	0.032	0.033
Dragon Touch	0.022	0.018	0.021	0.017	0.017	0.016	0.000	0.016
iPad 9.7	0.048	0.041	0.044	0.037	0.037	0.035	0.034	0.000

Table A5: **Estimated Diversion Ratios for Tablets (Plain Logit)**. Each cell reports the probability of choosing the row product j when the first choice, column product k , is removed from the choice set.

	Fire 7	Fire HD 8	Fire HD 10	Fire 7 Kids	iPad 10.2	Fire HD 8 Kids	Dragon Touch	iPad 9.7
Fire 7	0.000	0.590	0.162	0.362	0.143	0.033	0.024	0.080
Fire HD 8	0.753	0.000	0.429	0.336	0.200	0.193	0.108	0.116
Fire HD 10	0.058	0.144	0.000	0.079	0.142	0.297	0.022	0.077
Fire 7 Kids	0.127	0.131	0.080	0.000	0.012	0.440	0.293	0.007
iPad 10.2	0.049	0.079	0.141	0.010	0.000	0.015	0.310	0.715
Fire HD 8 Kids	0.003	0.034	0.151	0.200	0.004	0.000	0.188	0.001
Dragon Touch	0.000	0.001	0.000	0.011	0.013	0.019	0.000	0.003
iPad 9.7	0.011	0.020	0.036	0.002	0.485	0.003	0.055	0.000

Table A6: **Estimated Diversion Ratios for Tablets (Selected Model: *Descriptions ST*)**. Each cell reports the probability of choosing the row product j when the first choice, column product k , is removed from the choice set.