Adventures in Demand Analysis Using AI

Philipp Bach*,
Victor Chernozhukov†
Sven Klaassen‡
Martin Spindler§
Jan Teichert-Kluge¶
and
Suhas Vijaykumar‖

January 3, 2025

Abstract

This paper advances empirical demand analysis by integrating multimodal product representations derived from artificial intelligence (AI). Using a detailed dataset of toy cars on *Amazon.com*, we combine text descriptions, images, and tabular covariates to represent each product using transformer-based embedding models. These embeddings capture nuanced attributes, such as quality, branding, and visual characteristics, that traditional methods often struggle to summarize. Moreover, we fine-tune these embeddings for causal inference tasks. We show that the resulting embeddings substantially improve the predictive accuracy of sales ranks and prices and that they lead to more credible causal estimates of price elasticity. Notably, we uncover strong heterogeneity in price elasticity driven by these product-specific features. Our findings illustrate that AI-driven representations can enrich and modernize empirical demand analysis. The insights generated may also prove valuable for applied causal inference more broadly.

Keywords: Causal Inference; Deep Embeddings; Causal Fine-Tuning; Text; Images; Difference-in-Difference; Panel Data; Debiased Machine Learning; Demand Analysis; Foundation Models.

*University of Hamburg, philipp.bach@uni-hamburg.de
†Massachusetts Institute of Technology, vchern@mit.edu
‡University of Hamburg, sven.klaassen@uni-hamburg.de
§University of Hamburg, martin.spindler@uni-hamburg.de
¶University of Hamburg, jan.teichertkluge@uni-hamburg.de
‖Massachusetts Institute of Technology, suhasv@mit.edu

1 Introduction

Almost a century ago, *The Journal of the American Statistical Association* published early empirical studies of demand that applied statistical methods to measure how consumers respond to price changes: Work by scholars such as Wright (1929), Working (1943), Schultz (1933), Mills (1931, 1937a,b), and Stigler (1939) moved economics from theory towards quantitative measurement. By doing so, this work provided a foundation for econometrics, a field of statistical analysis focusing on economic problems. Their research established a tradition of using data to understand market behavior and inform economic models.

Today, advances in artificial intelligence (AI) and machine learning offer opportunities to build on this tradition. Instead of relying solely on simple numeric variables, researchers can now incorporate AI-generated product representations derived from text descriptions and images. These methods draw on hedonic modeling approaches (Griliches, 1971; Pakes, 2003) and integrate recent machine learning techniques (Devlin et al., 2019; Dosovitskiy et al., 2021), allowing economists to represent products more richly and capture nuances that standard covariates do not.

Using sales ranking and price data for toy cars on *Amazon.com*, we demonstrate how transformer-based models can leverage multiple, rich sources of product information for demand analysis. Our data include text descriptions, images, sales ranks, and prices. These multimodal inputs yield highly informative numerical embeddings that capture demand-relevant product attributes not easily summarized by standard human-encoded tabular variables—such as quality, branding, and visual characteristics—as we illustrate in Section 2.

We then fine-tune these embeddings to predict price and quantity signals, as these predictions are critical inputs to our causal inference problem. The resulting models achieve higher predictive accuracy than simpler specifications which rely solely on tabular data. Embeddings capture subtle distinctions between products—such as quality, branding, or visual characteristics—that influence

consumer demand and market prices, but are difficult to quantify using conventional methods. This improvement in predictive power suggests that AI-generated representations can meaningfully enhance empirical demand analysis and other causal inference tasks.

Finally, we address the challenge of estimating the price elasticity of demand, a central economic parameter. In our setting, simple cross-sectional regressions yield implausibly small elasticity estimates because they fail to capture product visibility and quality as key confounders. This motivates us to formulate a dynamic model with multimodal product attributes, along with lagged quantity and price signals, all of which serve both as confounders and as price-elasticity modifiers. By estimating such a dynamic model, we obtain more realistic price elasticities. Furthermore, we uncover pronounced heterogeneity in price elasticities that varies with product characteristics, as well as with how expensive and popular the products are. This underscores the economic value of AI-based representations: when properly fine-tuned, they yield more nuanced and credible estimates of how consumers respond to price changes across different products.

Our approach contributes to multiple strands of the literature. It extends empirical demand analysis by employing AI-generated, multimodal representations of products. Our work also builds on the emerging intersection of econometrics and machine learning (Athey and Imbens, 2019; Mullainathan and Spiess, 2017; Varian, 2014; Chernozhukov et al., 2018a) and complements recent studies that apply AI-based text analysis and other modern methods to economic questions (Belloni et al., 2014; Bajari et al., 2023; Compiani et al., 2023). In doing so, it provides a framework for combining flexible product representations with established econometric tools for identification and inference. It also introduces the idea that embeddings can and should be fine-tuned with causal inference in mind—which is critical in our context and potentially useful in other applications.

Our key empirical result is that AI-based embeddings are strong determinants/modifiers of

the price elasticity function. Interestingly, we also find that these embeddings are not major confounders of the relationship between price and quantity. Thus, while standard homogeneous price-effect models can yield ballpark estimates of average elasticity, they can severely understate or overstate the price elasticity for certain sets of products. These findings constitute important new empirical insights.

Our work also adds to a nascent literature in applied industrial organization that estimates demand models using public e-commerce data—such as ratings, reviews, or sales rankings—as proxies for the quantity sold, as the quantity sold is generally not published (Reimers and Waldfogel, 2021; He and Hollenbeck, 2020; Lee and Musolff, 2022). Building on He and Hollenbeck (2020) (see also Chevalier and Goolsbee, 2003), we treat the sales ranking as a proxy for relative quantity sold, motivated by the relationship among order statistics of the power law distribution. We show that combining this approach with high-quality embeddings of product descriptions and images produces realistic estimates of price elasticities.

The remainder of the paper is organized as follows: Section 2 discusses the use of AI-driven representations in demand analysis and describes the toy cars dataset, including how we extract and process multimodal features to create product embeddings. It also presents our first empirical results, highlighting how embeddings improve accuracy in predicting prices and quantities. Section 3 focuses on estimating the price elasticity; it lays out the underlying causal inference problem, discussing potential sources of confounding. We also illustrate the power of our embeddings in describing the heterogeneity of the price elasticity function. Finally, Section 4 concludes by summarizing our findings and discussing their implications for future research at the intersection of AI and econometrics. The Addendum contains deferred theoretical discussions, and the Online Appendix describes the workflow and algorithms used for feature generation and model estimation.

2 Using AI to Understand and Represent Products

2.1 The Data and Measurement of Prices and Quantities

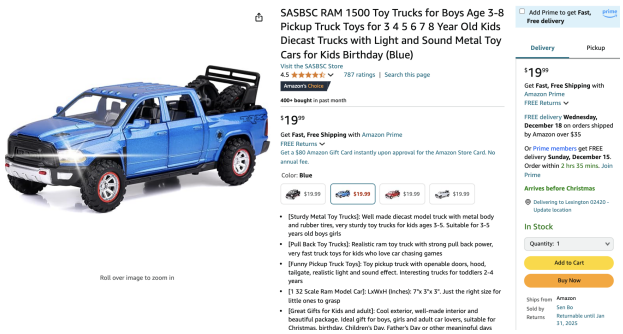

Figure 1: A product example with an image and text description in the "toys" category.

Our analysis uses a data set of toy cars from *Amazon.com*, compiled and provided by the data aggregator *Keepa.com*. For each item i, we collected its sales rank and price at time points spanning from April to December 2023. We also gathered each product's description, image, and additional tabular features (e.g., its subcategory on *Amazon.com*), as summarized in Table 1. Figure 1 illustrates a typical product page containing the product image and description. Overall, our data set comprises $N = 9{,}613$ unique products.

For our analysis, we define the quantity signal as

$$Q_{it} = \log\big(1/\text{Time-Averaged Sales Rank of } i \text{ in period } t\big),$$

and the price signal as

$$P_{it} = \log\big(\text{Time-Averaged Price of } i \text{ in period } t\big).$$

Each period, indexed by $t = 1,\ldots,T$, spans 4 weeks. We have $T = 9$ periods in total, each separated by 1 week. We structure our data set this way to limit inter-temporal feedback in our price elasticity analysis in Section 3. We also examine temporal changes in these signals, $\Delta Q_{it} := Q_{it} - Q_{i(t-1)}$ and $\Delta P_{it} := P_{it} - P_{i(t-1)}$.

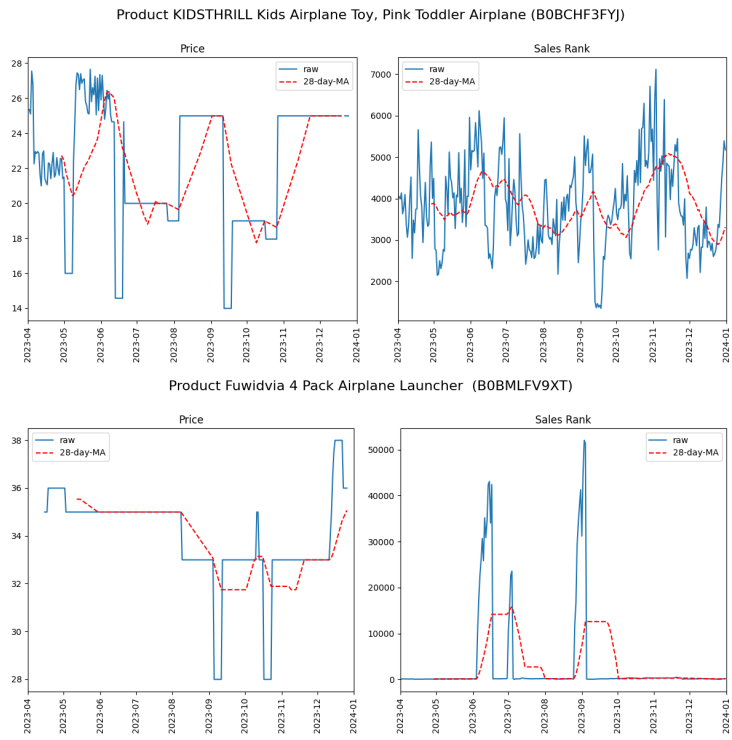

Figure 2: Price and sales rank series for two example products.

Note: The main series are shown as solid lines. The 28-day moving averages are shown in dashes.

In what follows, we refer to Q_{it} and P_{it} as the quantity and price signals, respectively. Under a power law assumption (see Remark 1 below), the logarithm of actual sales is proportional to the logarithm of the inverse sales rank; thus, we adopt inverse sales rank as our quantity signal. Furthermore, under this assumption, the price sensitivity of the inverse rank is proportional to the price sensitivity of the actual quantity sold, enabling us to capture demand responses to price changes.

Remark 1. Our use of the time-averaged inverse sales rank is an approximation motivated by modeling the latent, true quantities Q^*_{it} as independent draws from an underlying distribution at each time t. If an item has sales rank k, the quantity sold should be distributed as the k^{th} order statistic from a sample of size N. In the case of Pareto distributions with shape ϑ, a unit change in $\log Q^*_{it}$, $\Delta \log Q^*_{it}$, corresponds on average to a change of $-\vartheta \cdot \Delta \log Q^*_{it}$ in the log sales rank,

Table 1: Variable description, toy products data set.

Variable	Description
Sales rank	Ranking by units sold, relative to other products in the "games and toys" category
Price	Lowest offered price for a new product, excluding shipping and handling fees
Review count	Number of reviews
Rating	Average customer feedback rating
Lightning deals	Binary; 1 if product is promoted via a lightning deal
Buy box	Binary; 1 if the product delivery is fulfilled by Amazon
Product description	Unstructured text; includes title, manufacturer, brand, model, color, and size
Subcategory	Categorical; product subcategory
Product image	First image featured on the product page
ASIN	Unique product ID

Note: Product description, image, product ID, and subcategory do not vary with time in our data set; all other variables can vary with time.

$-Q_{it}$. This motivates our use of the sales rank. Using this approximation, He and Hollenbeck (2020) estimate $\hat{\vartheta} \approx 0.5$ for toys on *Amazon.com*; consequently, we can multiply our estimates by $1/\hat{\vartheta} \approx 2$ to obtain rough estimates of the price elasticity of demand. More generally, we can quantify the connection between our estimates and price elasticities under various assumptions about the distribution of sales (Office of National Statistics, 2020).

2.2 Using AI to Represent Products

To convert product data into useful numerical features, we employ various encoding models based upon the transformer neural network architecture proposed by Vaswani et al. (2017). We convert text descriptions into embeddings T_i using language models such as RoBERTa (Liu et al., 2019) or LLaMA 3 (Touvron et al., 2023), convert images into dense embeddings I_i using the BEiT model

(Bao et al., 2022), and transform tabular data into embeddings using the SAINT model (Somepalli et al., 2022). The Appendix provides additional implementation details.

While the aforementioned models are designed to work with the data we have on hand, the underlying approach is well suited for generalization to new types of data. In order to succeed in our context, these models must also be integrated and fine-tuned appropriately for estimating the price sensitivity. We discuss three key aspects of this approach which help explain its success and facilitate generalization to new contexts: self-supervised learning, the attention-based transformer architecture, and fine-tuning motivated by orthogonalized estimation of causal effects.

Self-Supervision. A significant challenge in machine learning is the scarcity of high-quality labeled data, as manual annotation is both expensive and time-consuming. Self-supervision addresses this limitation by creating labeled examples directly from unlabeled data. In this process, a portion of the input is deliberately masked or corrupted, and the model learns to predict these masked elements. This approach, fundamental to models like BERT (Devlin et al., 2019), effectively transforms each input sample into a self-labeled instance.

Consider the example sentence: $S =$ "Well made diecast model truck with metal body." We create a masked version: $W =$ "Well made [m] model truck with [m] body". The original sequence S serves as an auxiliary label that the model attempts to reconstruct from the corrupted input W. By applying this approach to billions of sentences, models learn to capture syntactic and semantic relationships without explicitly annotated labels.

The resulting internal representations, called *embeddings*, are extracted from the model's hidden layers and represent features of words or sentences. The approach generalizes well to other data types: for images, masking out patches and asking the model to predict the missing parts (He et al., 2022) enables the extraction of informative, context-dependent embeddings.

Attention. Transformer-based models employ so-called *attention mechanisms* (Vaswani et al.,

2017) to efficiently represent data; these underlie all of the embedding models used in this paper, as well as the highly influential GPT models (Brown et al., 2020). Attention refers to a specific structure repeated several times within transformer neural networks, which allows the model to selectively weight the most relevant components of the input when making predictions.

Through adaptive weighting of different input elements, attention produces embeddings that incorporate contextual and nuanced relationships between objects. This is believed to produce more informative embeddings than earlier, context-free models (such as word2vec or GloVe in the case of word embeddings; Mikolov et al., 2013; Pennington et al., 2014). Quantitatively, attention-based models achieve state-of-the-art performance across a wide range of tasks (Brown et al., 2020; Bao et al., 2022).

Causal Fine-Tuning. After a model has learned embeddings through self-supervision, it can be adapted for various downstream tasks. In our setting, the embeddings are important inputs to our causal inference problem: we are interested in how prices affect demand, holding fixed both product characteristics and other demand determinants. Our *fine-tuning* updates the pre-trained model parameters for this specific end-goal, by optimizing prediction of quantity and price signals. This is precisely the right target for orthogonal estimation of the price elasticity, as we further discuss in Section 3.4.

During fine-tuning, the embeddings serve as inputs to a specialized prediction layer. The errors from the prediction layers are then used to inform and update the parameters of the embeddings through gradient descent steps, which are computed via back-propagation (Rumelhart et al., 1986). The following diagram summarizes the process:

$$X_i^{in} = \begin{bmatrix} \text{Text}_i \\ \text{Image}_i \end{bmatrix} \xmapsto{e} E_i := \begin{bmatrix} T_i \\ I_i \end{bmatrix} \xmapsto{m} \{\hat{Q}_{it}, \hat{P}_{it}\}_{t=1}^{T}.$$

$$A_i^{tx} \updownarrow$$

$$A_i^{im} \updownarrow$$

Here, X_i^{in} represents the raw inputs (text Text_i and image Image_i). The embedding map e transforms these inputs into embeddings E_i. The terms A_i^{im} and A_i^{tx} are the auxiliary (masked) targets or "labels" in the self-supervised task. The model learns embeddings by attempting to reconstruct these targets from the unmasked parts of the inputs. The map m represents a downstream prediction layer that uses the embeddings E_i to predict tasks of interest, such as $\hat{Q}_{it}$ (quantities) and $\hat{P}_{it}$ (prices) over time.

As the model learns to reconstruct the auxiliary targets, it refines its internal representations. These improved embeddings E_i are then used by m to make high-quality predictions. Fine-tuning adjusts both e and m to ensure that the embeddings and downstream predictions align with the target predictive and causal inference questions.

2.3 Evaluating the Embeddings

After obtaining the embeddings, we must assess whether they effectively represent the products and "understand" their characteristics. We first take the concatenated embeddings $E_i = \left(T_i, I_i\right)$, where T_i also includes tabular embeddings, and then apply a Johnson–Lindenstrauss projection of these embeddings onto a 256-dimensional vector $\bar{E}_i$; Johnson (1984). This projection approximately preserves distances and is therefore considered (at least approximately) information-lossless. We

then center and normalize the embeddings so they lie on a hypersphere:

$$X_i^e := \frac{\bar{E}_i - \frac{1}{n}\sum_i \bar{E}_i}{\left\| \bar{E}_i - \frac{1}{n}\sum_i \bar{E}_i \right\|}.$$

We use these normalized embeddings in our subsequent analysis.

We evaluate these embeddings through two approaches:

1. **Qualitative.** We examine similar products or clusters of products on this hypersphere and assess the results qualitatively.

2. **Quantitative.** We determine whether these AI-generated features improve predictions of price and quantity signals, where predictions serve as key inputs into downstream causal inference.

Both approaches are crucial for demand analysis, including the computation of hedonic inflation prices, forecasting demand and prices for new products, and understanding how demand responds to price variations.

2.3.1 Qualitative Assessment

For the clustering task, we perform k-means clustering to group products into five clusters based on their embeddings. To examine the influence of images, we first cluster using both text and image embeddings, and then using only text embeddings. We visualize the resulting product clusters in three-dimensional space by projecting the embeddings onto the first three principal components, as shown in Figures 3 and 4.

When text and image embeddings are combined, the projection yields a "full" ball of product points with distinctly separated clusters. In contrast, using text-only embeddings produces a "stripe on a sphere," where the points are concentrated near the boundary and around the equator of the ball. Nevertheless, the clusters remain well-separated even without the image information.

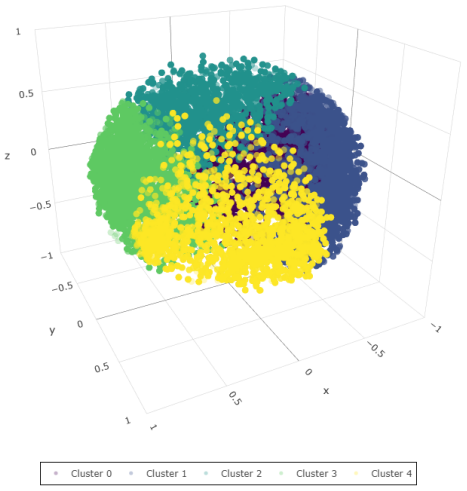

Figure 3: 3d representation of product embeddings (with image) and five clusters

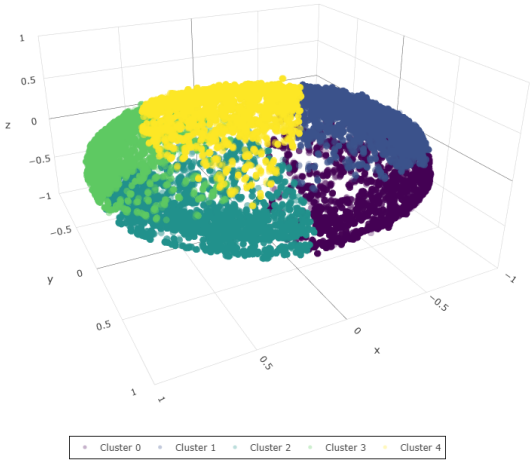

Figure 4: 3d representation of product embeddings (no image) and five clusters

While mainly illustrative, these visuals suggest that text-only embeddings lie in a lower-dimensional space, missing valuable image-based information absent from text descriptions. This observation is supported by Tables 2 and 3, which show that clusters formed from text+image embeddings have more visually coherent centroids and exhibit greater internal homogeneity—confirming the importance of multimodal data.

To further explore these clusters, we employ generative AI tools to summarize and characterize each cluster centroid. We also construct an "average" representative image for products near the centroids, with results shown in Table 4. These results closely match our own assessments of the product cluster centers, underscoring the utility of combining both text and image embeddings for cluster analysis.

Cluster 0	 Bruder Land Rover Fire Department Vehicle wit...	 Carrera 27625 Porsche 935 GT2 No. 96/69 1:32 ...	 Bruder 02509 RAM 2500 Service Truck with Rota...
Cluster 2	 Vokodo Friction Powered Toy Semi Truck Traile...	 Anti-Reverse Car Toy Set, 2022 New Reverse Ca...	 Srenta 4 inch Die Cast Pull-Back Monster Trac...
Cluster 4	 Movie Car Toys Car 1:55 Diecast Vehicles for ...	 Movie Car Toys Car 1:55 Diecast Vehicles for ...	 Movie Car Toys Car 1:55 Diecast Vehicles for ...

Table 2: Examples of closest products to cluster centers (Tabular + Text + Image); Only three examples of clusters shown.

2.3.2 Quantitative Assessment

While the previous discussion provides a qualitative indication that the model can represent products effectively, we now present a more quantitative assessment of the model's predictive performance.

We begin by examining how well the embeddings—and their "compressed" versions—predict price and quantity levels, as well as their changes. Formally, our targets are $Y \in \{Q, P, \Delta Q, \Delta P\}$. We also use these predictive regressions to fine-tune the embeddings themselves.

As shown in Table 5, simple linear regressions using only tabular data perform poorly. Boosted trees yield substantial gains in predictive accuracy, and neural networks with text embeddings perform even better. Including image embeddings offers further improvements, though the additional gains are modest (e.g., a 1.5 percentage-point increase in R^2 for Q_{it}). Nonetheless, these gains are meaningful.

We also assess each model's ability to predict changes in quantities and prices, rather than their levels. As expected, predicting changes is notably more difficult, leading to a sharp decline

Cluster 0	 FFTROC Dinosaur Toys for Kids 3-5 Pull Back C...	 Hot Wheels GBG26 Mario Kart 1:64 Die-Cast Mar...	 Hot Wheels GBG28 Mario Kart 1:64 Die-Cast Pea...
Cluster 2	 Doloowee Construction Truck Toys for Toddlers...	 Doloowee Truck Toys for Toddlers 3 4 5 6 Year...	 Monster Jam, Champ Ramp Freestyle Playset wit...
Cluster 4	 Gemini&Genius Construction Vehicle Toys Timbe...	 Haomabu Lexus LX570 1/32 Off-Road SUV Car Mod...	 Haomabu Lexus LX570 1/32 Off-Road SUV Car Mod...

Table 3: Examples of closest products to cluster centers (Tabular + Text); Only three examples of clusters shown.

in performance. Even the best models achieve only about 18% R^2 for changes in quantity and 1% R^2 for changes in price. Despite this drop, the results underscore that AI-generated features provide strong predictive power for quantity levels and some improvement—albeit smaller—for predicting changes.

Next, we investigate whether "compressed" embeddings preserve the information in the full-dimensional embeddings. Specifically, we consider:

- **Principal Components (PCA):**

$$X_{i,k}^{pc} := \gamma_k^T X_i^e, \quad X_i^{pc} := (X_{i,k}^{pc})_{k=1}^K,$$

  where γ_k is the k-th eigenvector of the covariance matrix of X_i^e, corresponding to the k-th largest eigenvalue.

- **Centroid Similarities (CS):**

$$X_{i,k}^{sim} := c_k^T X_i^e, \quad X_i^{sim} := (X_{i,k}^{sim})_{k=1}^K,$$

  where c_k is the centroid (mean) of the k-th cluster identified by k-means.

Cluster 0	Premium-Detail Replicas of Emergency, Service, and Racing Vehicles for Realistic Play	
Cluster 1	Oversized and Interactive Truck Playsets Offering Transforming Functions and Action-Packed Fun	
Cluster 2	Pull-Back, Friction-Powered, and Easily Transforming Vehicles for Fast, Engaging Adventures	
Cluster 3	A Variety of Diecast Classics and Specialty Vehicles for Collectors and Everyday Play	
Cluster 4	Iconic Movie-Inspired 1:55 Scale Diecast Cars Perfect for Storytelling and Roleplay	

Table 4: Generative AI summaries and images for the five cluster centroids (Tabular + Text + Image).

These vectors capture how similar the embedding vectors are to principal axes of variation—either principal components or k-means cluster centroids—using cosine similarity. (E.g., since c_k and X_i^e lie on the unit hypersphere, their inner product $c_k^\top X_i^e$ directly represents the cosine similarity). As shown in Table 6, using only five principal components or five centroid similarities can retain nearly all the prediction-relevant information contained in the original embeddings. In particular, when using a boosted tree, these compressed embeddings nearly match the performance of a deep neural network that uses the entire text and image inputs. This finding can simplify downstream tasks. In this paper, we rely primarily on centroid similarities, as they appear more interpretable than principal components for our application.

Overall, these results suggest that AI-derived embeddings greatly improve predictions of price and quantity levels, although they are less effective for predicting price changes. This has crucial implications for causal (price sensitivity) analysis, discussed below. Notably, our findings indicate

Table 5: Test R^2 scores for predicting quantity and price signals.

Method [features] \Target	Q_{it}	P_{it}	ΔQ_{it}	ΔP_{it}
Linear Reg [all tabular]	22.08%	17.10%	6.11%	0.36%
Boosted Trees Reg [all tabular]	45.84%	20.37%	13.21%	-2.60%
Deep Learning Reg [text only; invariant tabular]	50.99%	63.12%	0.01%	-0.00%
Deep Learning Reg [image and text; invariant tabular]	51.24%	64.67%	0.05%	-0.07%
Deep Learning Reg [text only; all tabular]	60.04%	65.15%	18.41%	0.76%
Deep Learning Reg [image and text; all tabular]	61.77%	65.43%	11.35%	1.03%

Note: All models are trained on a training set and scores are evaluated on a test set. Predictions use lagged values of time-varying controls.

that product embeddings tend to act more as *effect modifiers* (i.e., determinants of elasticity) rather than *confounders* of the causal relationship.

3 Estimating Price Effects

Understanding how price changes affect consumers' choices is a central challenge in empirical economics and marketing. One common way to measure this relationship is through the elasticity of demand with respect to price. Although a regression of sales on prices may appear a straightforward way to estimate elasticity, it can yield biased estimates if key confounding factors are not properly accounted for. In this section, we explore various approaches to uncover the true causal price sensitivity and discuss their respective strengths and limitations.

Table 6: Test R^2 Scores for ML methods using DL-based PCAs and similarities together with tabular controls.

Method [+ DL Features] \Target	Q_{it}	P_{it}	ΔQ_{it}	ΔP_{it}
Linear Reg [+5 PCAs]	48.69%	59.07%	6.15%	0.40%
Linear Reg [+ 5 Similarities]	48.60%	57.70%	6.16%	0.43%
Linear Reg [+256 Embeddings]	51.08%	59.66%	5.81%	-0.04%
Boosted Trees Reg [+5 PCAs]	56.02%	63.17%	14.99%	-1.28%
Boosted Trees Reg [+5 Similarities]	56.12%	62.28%	15.37%	-0.30%
Boosted Trees Reg [+256 Embeddings]	55.73%	64.05%	14.03%	-0.72%

Note: All models are trained on a training set and scores are evaluated on a test set. Predictions use lagged values of time-varying controls.

3.1 Initial Approach and Challenges

A natural starting point is to estimate the relationship between price and product performance using a predictive model. Consider a regression of the (log) inverse ranking of product i at time t, denoted Q_{it}, on the (log) price P_{it} and a set of controls $X_{it} = (X_i^e, X_{it}^o)$, where X_{it}^o represents other tabular controls:

$$\mathrm{L}[Q_{it} \mid P_{it}, X_{it}] = \delta P_{it} + g_t(X_{it}),$$

where $g_t(\cdot)$ is a function describing how the control variables influence the outcome over time. We allow g_t to vary with t. The notation $\mathrm{L}[Y \mid P, X]$ denotes the projection of the random variable Y onto the space of partially linear prediction rules of the form $aP + g(X)$.

In practice, directly estimating this model often suggests a very small price sensitivity (or "elasticity"), captured by the coefficient δ: $\delta \approx [0, -0.2]$. From a causal perspective, this result is implausible: the notion that a price change exerts virtually no effect on ranking or sales is

counterintuitive, as it implies that raising prices would only marginally reduce the quantity sold. In other words, this finding suggests that certain key confounders—such as latent quality or visibility—are not adequately controlled for, resulting in a strongly biased estimate.

3.2 The Causal Dynamic Model and Regressions

To address the issue above, we introduce a simple dynamic panel data model to guide our statistical analysis. We can view the outcomes and key variables as arising from the following structural equation model (SEM):

$$Q_{it} = a_t\big(S_{it}, \varepsilon_{it}\big)\, P_{it} + q_t\big(S_{it}, \varepsilon_{it}\big), \tag{1}$$

$$P_{it} = p_t\big(S_{it}, \varepsilon^p_{it}\big), \tag{2}$$

$$S_{it} = s_t\big(S_{i,t-1}, \varepsilon^s_{it}\big); \quad S_{it} \equiv (Q_{i,t-1}, P_{i,t-1}, X_{it}), \tag{3}$$

where a_t, q_t, p_t, and s_t are nonparametric structural functions, and ε_{it}, ε^p_{it}, and ε^s_{it} are i.i.d. stochastic vectors that are mutually independent.

This specification defines an *autoregressive* model in which the quantity signal Q_{it} depends on the price signal P_{it} and other state variables S_{it}. The state variables include lagged quantity and price, $Q_{i,t-1}$ and $P_{i,t-1}$, time-invariant product characteristics X_i (captured through embeddings), and time-varying characteristics X^o_{it} (such as ratings and the number of reviews). Among these variables, the lagged quantity $Q_{i,t-1}$ is arguably a key confounder, reflecting both product visibility and quality—a conclusion reinforced by our empirical findings below. In particular, including the lagged quantity in the model substantially shifts the estimated price elasticity into a more plausible range.

Because the model follows a Markovian structure, each period updates the state variables, after which prices and quantities respond to the new state vector. Figure 5 illustrates this SEM

graphically.

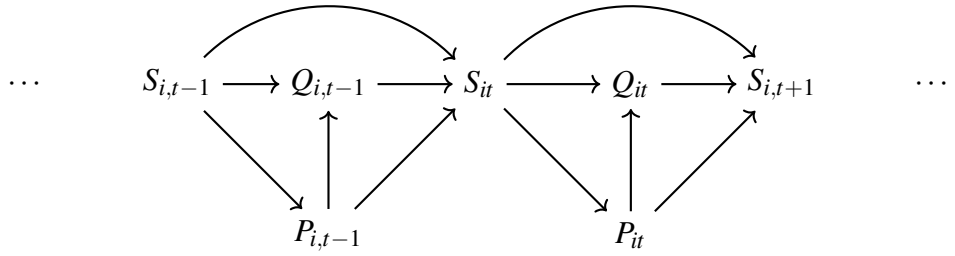

Figure 5: A directed acyclic graph for the dynamic model.

It is convenient to define the variable $A_{it} := a_t(S_{it}, \varepsilon_{it})$, which we interpret as a random elasticity or price sensitivity. Under this notation, the SEM above induces the potential outcomes (Rubin, 1975)

$$Q_{it}(p) = A_{it}\, p + q_t(S_{it}, \varepsilon_{it}),$$

by setting $P_{it} = p$ in the first equation; see Pearl (1995). Hence, the price sensitivity A_{it} is the causal effect of increasing P_{it} by one unit: $\partial_p Q_{it}(p) = A_{it}$. We focus on either the Average Causal Effect (ACE):

$$\alpha_t = \mathrm{E}\big[\partial_p Q_{it}(p)\big] = \mathrm{E}[A_{it}],$$

or the Conditional ACE (CACE):

$$\alpha_t(S_{it}) = \mathrm{E}\big[\partial_p Q_{it}(p) \mid S_{it}\big] = \mathrm{E}[A_{it} \mid S_{it}],$$

which describes the average causal effect conditional on product characteristics and thus captures the predictable component of price sensitivity.

Identification of both the CACE and the ACE in this setting follows from computing the conditional expectation of Q_{it} while conditioning on P_{it} (the treatment) and S_{it} (the observed confounders). Indeed, conditioning on S_{it} blocks non-causal sources of association between the outcome and the treatment (Pearl, 1995). Including further lags of state variables S_{it} is not strictly

necessary under the Markovian structure, but it can serve as a useful specification check. Below, we also discuss potential threats from unobserved confounders, such as time-varying demand shocks that may affect both the outcome and the treatment.

We now derive the key regression function of interest:

$$\mathrm{E}\big[Q_{it} \mid P_{it}, S_{it}\big] = \alpha_t(S_{it})\, P_{it} + \gamma_t(S_{it}), \tag{4}$$

where $\gamma_t(S_{it}) := \mathrm{E}\big[q_t\big(S_{it}, \varepsilon_{it}\big)\big]$, so the CACE function $\alpha_t(S_{it})$ appears as the heterogeneous slope in (4). The ACE parameter α_t then follows by averaging the CACE function over S_{it}.

3.3 Empirical Models

In the empirical analysis, we examine two forms of the CACE function:

$$\textbf{I. Homogeneous Effect:} \qquad \alpha_t(S_{it}) = \alpha_t; \tag{5}$$

$$\textbf{II. Heterogeneous Effect:} \qquad \alpha_t(S_{it}) = a_{0t} + \sum_{k=1}^{K} \alpha_{kt}\, X_{i,k}^{sim} \;+\; b_{1t}\, P_{i,t-1} \;+\; b_{2t}\, Q_{i,t-1}. \tag{6}$$

The first specification is very simple and serves as our baseline. The second is more elaborate yet still structured, allowing the elasticity function $s \mapsto \alpha_t(s)$ to depend on product characteristics as well as past quantities and prices:

- The first component of $\alpha_t(s)$ captures product characteristics in the product space, represented by similarity vectors describing the product's position.
- The second part lets the elasticity vary with how popular the products are (lagged quantity) and how expensive they are (lagged price).

We show empirically that both components matter. In presenting our results, we assume time homogeneity by setting $\alpha_t(\cdot) = \alpha(\cdot)$. Empirically, this did not affect any findings; we adopt this simplification purely for clarity of presentation.

For the model of the "control" function $\gamma_t(S_t)$, we consider three cases:

1. **Linear in State** S_{it}: $\gamma_t(S_{it}) = g_t^T S_{it}$.

2. **Interactive Linear in State** S_{it}: $\gamma_t(S_{it}) = d_t^T I(S_{it})$, where $I(S_{it})$ includes S_{it} and interactions of $P_{i,t-1}$ and $Q_{i,t-1}$ with X_{it}^{sim}.

3. **Nonlinear in State** S_{it}: $\gamma_t(S_{it})$ is approximated by boosted trees.

The final model is fully nonparametric. We also experiment with using similarity vectors X_i^{sim} in place of the full 256-dimensional embedding X_i^e as controls, and find that the similarity vectors perform comparably well.

In summary, we will consider six types of empirical models, formed by the Cartesian product $\{\text{I},\text{II}\} \times \{1,2,3\}$. Within each of these six types, we also vary how the control variables are included. As a preview of results, we note that the heterogeneous-effects model (II) receives the strongest empirical support, with variants of types 1–3 yielding similar quantitative results on elasticity.

3.4 Orthogonal Inference of Causal Effects

We identify and estimate the causal effects using the following projection equation:

$$Q_{it}^{\perp} = \delta_t(S_{it}) P_{it}^{\perp} + e_{it}, \quad e_{it} \perp P_{it}^{\perp} \mid S_{it}, \tag{7}$$

where $e_{it} \perp P_{it}^{\perp} \mid S_{it}$ means $\mathrm{E}[e_{it} P_{it}^{\perp} \mid S_{it}] = 0$. The pair $\left(Q_{it}^{\perp}, P_{it}^{\perp}\right)$ consists of the residuals

$$Q_{it}^{\perp} = Q_{it} - \mathrm{E}[Q_{it} \mid S_{it}], \quad P_{it}^{\perp} = P_{it} - \mathrm{E}[P_{it} \mid S_{it}].$$

The coefficient function $\delta_t(S_{it})$ is the conditional predictive effect (CAPE) of a shock in the exposure variable on a shock in the outcome:

$$\delta_t(S_{it}) := \frac{\mathrm{E}\left[Q_{it}^{\perp} P_{it}^{\perp} \mid S_{it}\right]}{\mathrm{E}\left[P_{it}^{\perp 2} \mid S_{it}\right]}.$$

Averaging the CAPE gives the average predictive effect (APE), $\mathrm{E}[\delta_t(S_{it})]$.

We now make the following observation.

Proposition 1 (Identification of CACE)**.** *If the SEM model* (1)–(3) *holds, then the CACE is identified by the CAPE:*

$$\alpha_t(S_{it}) = \delta_t(S_{it}),$$

almost surely, provided that both exist and are finite. Then the ACE is identified by the APE, $\mathrm{E}\big[\alpha_t(S_{it})\big] = \mathrm{E}\big[\delta_t(S_{it})\big]$*, again provided these expectations exist and are finite.*

This claim follows directly from the regression equation (4), by defining

$$e_{it} := q_t\big(S_{it}, \varepsilon_{it}\big) \; - \; \gamma_t\big(S_{it}\big),$$

and then verifying the required orthogonality between e_{it} and $P_{it}^{\perp}$.

Once we learn the CAPE, we effectively learn the CACE, provided that the causal SEM postulated above holds. If the SEM is only approximate, then the CAPE can still be treated as an approximation to the CACE; we elaborate on this in Section 3.6.

We estimate the CAPE and CACE using both linear regression and nonlinear, nonparametric models, applying modern machine-learning tools and cross-fitting to compute the residualized outcomes and exposure variables. We then estimate the projection equation (7) using either homogeneous or heterogeneous forms of $\delta_t(\cdot) = \alpha_t(\cdot)$ via least squares, and apply conventional statistical inference to construct p-values and confidence intervals, following Chernozhukov et al. (2018a). Details on the workflow and implementation algorithms are provided in the Online Appendix.

Remark 2 (Orthogonalization). The argument above relies on the classical partialling-out or orthogonalization approach (Frisch and Waugh, 1933; Lovell, 1963; Robinson, 1988). The "residual-on-residual" method underlies double machine learning (also called *R*-learning), which

uses cross-fitted machine learning to estimate residuals and then infers the CAPE by least squares (Chernozhukov et al., 2018a; Nie and Wager, 2021; Semenova et al., 2023). This approach is part of a broader class of debiased machine learning (DML) algorithms rooted in semiparametric learning theory (Levit, 1975; Hasminskii and Ibragimov, 1978; Pfanzagl and Wefelmeyer, 1985). □

Remark 3 (Causal Fine-Tuning). Recall that we fine-tuned the embeddings X_i^e to produce the best-performing prediction rules for Q_{it} and P_{it} (or their temporal differences, since past values strongly predict future outcomes). In this sense, the fine-tuning was well-suited to our causal inference problem. This idea generalizes to fully nonlinear models, where one could fine-tune embeddings to learn the Neyman-orthogonal equations for the parameter of interest. □

Remark 4 (Robustness of DML to Estimation Noise in Embeddings). One might suspect that using estimated embeddings rather than "optimal" ones would complicate inference. However, under mild conditions, this is not the case. The key defining property of DML is that its estimating equations are robust to perturbations in the nuisance regression function—a feature referred to as Neyman orthogonality. Since perturbations in regressors translate to perturbations in the regression function, the former produce a *zero* first-order effect on the DML estimator. We provide further theoretical details on this point in Addendum A. □

3.5 Empirical Results

3.5.1 Homogeneous Elasticity Model

We begin by examining the homogeneous elasticity function. Table 7 shows that, in various specifications—ranging from simpler linear regressions with lagged price and quantity to more complex boosted-tree models incorporating product embeddings and additional controls—price consistently exerts a negative and highly significant effect on the quantity signal (the inverse sales rank). The confidence intervals range from about -0.6 to -0.45, indicating strong economic and

Table 7: Estimated price effects based on the partially linear dynamic model.

Specification of Control Function (State S_t)	coef	std err	t	P-val.	[5.0%	95.0%]
I-1. Linear (P_{t-1}, Q_{t-1})	-0.542	0.041	-13.372	0.000	-0.608	-0.475
I-1. Linear (P_{t-1}, Q_{t-1}, X^e,X_t^o)	-0.528	0.040	-13.208	0.000	-0.594	-0.463
I-1. Linear (P_{t-1}, Q_{t-1}, X^{sim},X_t^o)	-0.527	0.040	-13.140	0.000	-0.593	-0.461
I-2. Linear with Interactions (P_{t-1}, Q_{t-1}, X^{sim},X^o)	-0.529	0.040	-13.326	0.000	-0.595	-0.464
I-3. Boosted Trees (P_{t-1}, Q_{t-1}, X^e,X_t^o)	-0.538	0.038	-14.041	0.000	-0.601	-0.475
I-3. Boosted Trees (P_{t-1}, Q_{t-1}, X^{sim},X_t^o)	-0.513	0.039	-13.191	0.000	-0.577	-0.449

Note: Standard errors are clustered at the product level. The LR models are estimated using OLS. The PLR model is estimated using DML with cross-fitted boosted trees.

statistical significance. Indeed, these estimates imply that a 1% price increase reduces the inverse sales rank by about $[0.6, 0.45]$%. Furthermore, to convert this effect into an elasticity of the actual demand under the power law assumption (cf. Remark 1), we need to multiply these coefficients by about 2, which indicates that the demand elasticity itself falls in the range of approximately $[-1.2, -0.9]$.

The results consistently suggest that incorporating lagged quantities and prices—which capture a product's popularity and quality—reveals a more substantial, economically meaningful negative price elasticity. Interestingly, other product characteristics, including embeddings and similarities, are not statistically or economically significant confounders here. This finding is consistent with the results in Section 2, which indicate that product embeddings or similarities do not strongly predict price changes. Hence, including or excluding these controls does not substantially alter the estimates. However, this does not rule out the possibility that product embeddings could be crucial *effect modifiers*, as we discuss next.

3.5.2 Heterogeneous Elasticity Model

While the homogeneous specification yields more plausible estimates of average price sensitivity, it may still obscure important differences across products. The next step, therefore, models heterogeneous effects, allowing price elasticity to vary as a function of observed product characteristics or embedding-based similarity measures. Figure 6 and Table 8 illustrate how the elasticity estimates relate to these cluster similarities and to lagged price and quantity signals. These results reveal that certain products are more price-sensitive than others. For example, higher-priced and better-ranked products show greater price sensitivity, and similarities (to cluster centroids) also emerge as key drivers of sensitivity. (The first point is perhaps clearer from Table 8.) Products associated with specific clusters experience notably different price-sensitivity levels, underscoring the importance of accounting for product-level heterogeneity when estimating price effects.

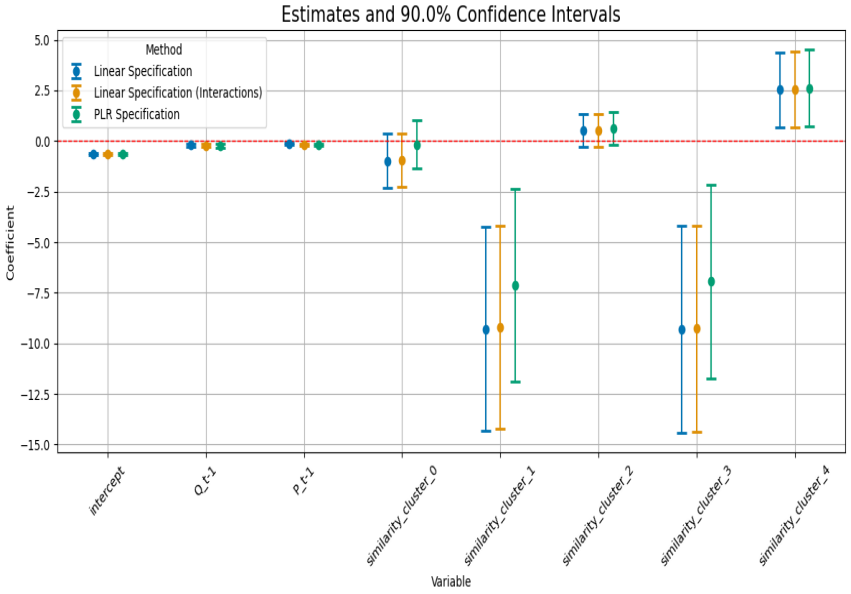

Figure 6: Parameters of the Estimated Elasticity Function and 90% Confidence Intervals

Note: Covariance is clustered at the product level. The LR model is estimated using OLS. The PLR model is estimated using DML with cross-fitted boosted trees.

To provide a comprehensive view of how strong this heterogeneity can be, Figure 7 plots

Table 8: Inference on the price effect modifiers with nonlinear (Boosted Trees) state control (Model II-3). Results for Models II-1 and II-2 are similar.

Modifier	coef	std err	t	P-val.	[5.0%	95.0%]
Centercept	-0.643	0.039	-16.643	0.000	-0.706	-0.579
Lagged Quantity, Q_{t-1}	-0.226	0.056	-4.035	0.000	-0.318	-0.134
Lagged Price, P_{t-1}	-0.167	0.031	-5.322	0.000	-0.219	-0.115
Cluster Similarity 0	-0.179	0.722	-0.248	0.804	-1.367	1.009
Cluster Similarity 1	-7.134	2.890	-2.469	0.014	-11.887	-2.380
Cluster Similarity 2	0.630	0.499	1.261	0.207	-0.191	1.451
Cluster Similarity 3	-6.932	2.905	-2.386	0.017	-11.710	-2.154
Cluster Similarity 4	2.615	1.158	2.259	0.024	0.711	4.519

Note: Standard errors are clustered at the product level. The LR model is estimated using OLS. The PLR model is estimated using DML with cross-fitted boosted trees. Lagged quantities and prices are centered and rescaled to have unit variance across (i,t). The coefficient on the centercept represents the average effect.

each product's estimated elasticity (vertical axis) sorted in ascending order (horizontal axis is the "percentile index"); as per Chernozhukov et al. (2018b). The solid blue line represents the point estimates, the blue band denotes the 90% pointwise confidence intervals, and the dashed red line indicates the overall average elasticity. Reflecting the discussion of heterogeneous price sensitivity, elasticity ranges from about -1.4 to nearly 0, and these wide differences are statistically significant (as shown by the confidence bands). By multiplying these numbers by 2, we estimate that actual demand elasticity spans roughly -2.8 to 0, with an average near -2.1. This range highlights the importance of allowing for heterogeneity in price responses, as some products appear far more (or far less) sensitive to price changes than the average.

Tests of joint significance further confirm that cluster similarities matter. Table 9 presents p-values from a χ^2-test that evaluates whether the cluster similarity measures jointly have a

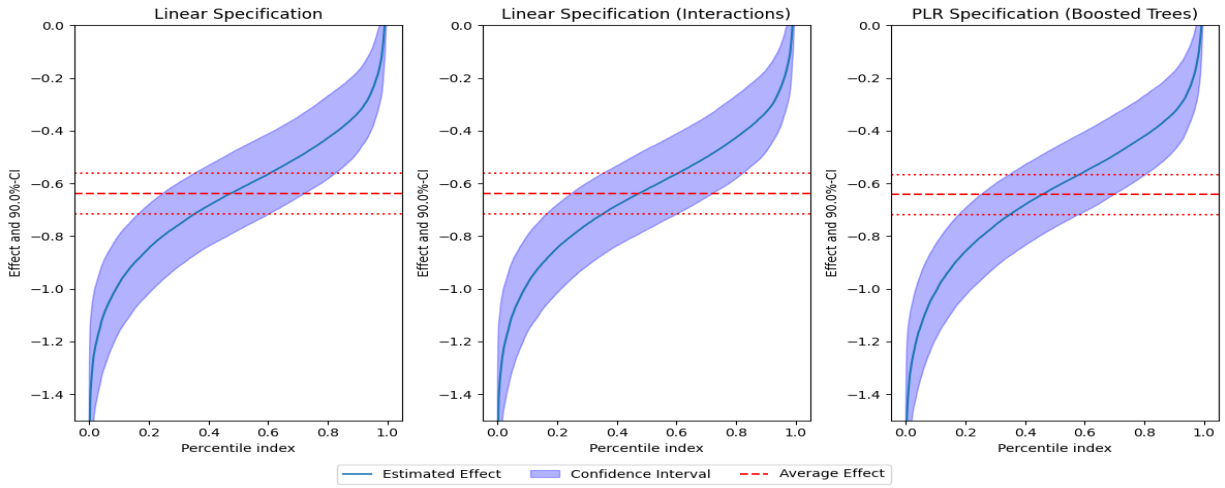

Figure 7: Sorted Elasticity as a Function of Effect Modifiers and 90% Pointwise Confidence Bands

Note: Covariance is clustered at the product level. The LR model is estimated using OLS. The PLR model is estimated using DML with cross-fitted boosted trees.

statistically significant impact on price sensitivity. The results indicate that these cluster-based heterogeneities are significant, reinforcing the conclusion that certain products are more responsive to price changes than others.

Table 9: p-values for χ^2-test of joint significance of price effect modifiers

Model	All Modifiers	Similarities Only
II.1 Linear Specification	0.000	0.031
II.2 Linear Specification (Interactions)	0.000	0.031
II.3 PLR Specification (Boosted Trees)	0.000	0.077

Note: Standard errors are clustered at the product level. LR is estimated using OLS. The PLR model is estimated using DML with cross-fitted boosted trees (sample size: $24,895$).

3.6 Some Reflections on the Limitations of the Analysis

Our statistical estimates carry a well-defined causal interpretation under the stated, relatively strong assumptions. The main threat to this interpretation is the presence of latent, time-varying factors that bias the relationship between P_{it} (the price signal) and Q_{it} (the quantity signal). Indeed, we might suspect that price endogenously responds to demand shocks ε_{it} from the outcome equation, effectively making ε_{it} a confounder. The DAG below illustrates such a scenario.

However, in our setting, we suspect that the link $\varepsilon_{it} \to P_{it}$ may be weak, as prices often follow "sticky," piecewise-constant paths that do not change as frequently as quantity signals (sales ranks); see, for example, Figures 2 in Section 2. Formally, if the edge from ε_{it} to P_{it} is zero, our earlier identification strategy holds, and our estimates are indeed causal. Otherwise, we can treat them as approximations of the causal estimates. For methods to bound the effect distortion caused when ε_{it} affects P_{it}, see Chernozhukov et al. (2021).

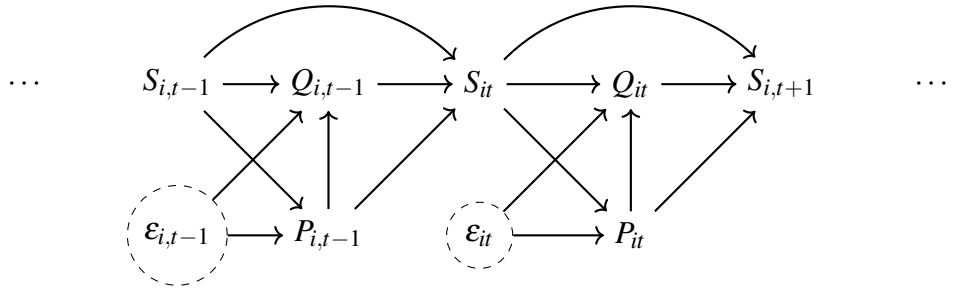

Figure 8: Dynamic model with demand shocks as confounders.

Another way to achieve identification is through an instrumental variable Z_{it} that induces exogenous variation in P_{it}, independent of ε_{it}. By conditioning on the variation produced by Z_{it}, we isolate a component of P_{it} uncorrelated with ε_{it}, which then identifies the causal effect of price changes on quantity changes. Essentially, one can estimate the average causal effect of Z_{it} on Q_{it} (controlling for S_{it}), and the average causal effect of Z_{it} on ΔP_{it} (also controlling for S_{it}), and then take their ratio—following the classical approach of Philip Wright (1928), republished

as Wright (2024)—to identify the causal effect of P_{it} on Q_{it}. We could not find such credible instruments in our setting. Using further lagged values of $Q_{i,t-1}$ is one potential approach, but it would require taking the first-order autoregressive specification for outcome too literally (i.e., those further lagged variables should likely be included as state variables, rather than employed as instruments).

4 Concluding Remarks

This study highlights the significant potential of AI-generated multimodal embeddings in demand analysis. By integrating text, image, and tabular data into a causal econometric framework, we improve both the precision of demand and price forecasts and the credibility of elasticity estimates. Our findings show that these rich embeddings not only enhance predictive accuracy but also reveal substantial heterogeneity in consumer price sensitivity, providing nuanced insights into demand behavior. These advancements create a methodological bridge between machine learning and econometrics, illustrating how modern AI tools can enrich traditional economic analyses. In future research, we hope to further explore the intersection of AI and causal inference, including bias-bounding and instrumental variable strategies to address time-varying confounding.

A Addendum: Robustness to Estimated Embeddings

One may suspect that using estimated embeddings instead of "optimal" ones could complicate inference. However, under mild conditions, this is not the case. To illustrate this point in a simple manner, consider the homogeneous model:

$$Q_{it}^{\perp} = \delta_t P_{it}^{\perp} + e_{it}, \quad e_{it} \perp P_{it} \mid S_{it}.$$

Let Y_{it} denote the predictive target, which is either P_{it} or Q_{it}. Let $X_i^e(\hat{\phi})$ denote the estimated

and fine-tuned embeddings, where $\hat{\phi}$ denotes estimated parameters, and $S_{it}(\hat{\phi})$ the derived controls. We assume that $\hat{\phi}$ is obtained from data that are independent of the main data used in the analysis. Similarly, let X_i^e and S_{it} denote the *ideal* embeddings and controls, in the sense that

$$\mathrm{E}[Y_{it} \mid S_{it}, S_{it}(\hat{\phi})] = \mathrm{E}[Y_{it} \mid S_{it}].$$

In other words, after including S_{it}, the best prediction rule for Y_{it} given both S_{it} and $S_{it}(\hat{\phi})$ depends only on S_{it}.

Consider a learner $\hat{\gamma}_t^Y(S_{it}(\hat{\phi}))$ that minimizes empirical risk $\sum_{i\in A}\{Y_{it} - \gamma(S_{it}(\hat{\phi}))\}^2$ over control functions γ in the convex model $\mathscr{F}$, conditional upon fine-tuned embeddings $\hat{\phi}$ (this includes our considered models 1-3). Here A is a subset of $\{1,..,n\}$ whose size is proportional to n. The DML inference approaches using ideal and estimated embeddings are first-order equivalent under the following two key conditions:

(E1) For the given $\hat{\phi}$, the *square root* of the offset Rademacher complexity (Liang et al., 2015) of the class $\mathscr{F}_{\hat{\phi}} = \{\gamma(S_{it}(\hat{\phi})) : \gamma \in \mathscr{F}\}$ is $o_p(n^{-1/4})$.

(E2) The approximation error of the model $\mathscr{F}$ with estimated embeddings $\hat{\phi}$ is sufficiently small: $\inf_{\gamma\in\mathscr{F}} \sqrt{\mathrm{E}[\{\mathrm{E}[Y_{it} \mid S_{it}] - \gamma(S_{it}(\hat{\phi}))\}^2 \mid \hat{\phi}]} = o_p(n^{-1/4})$.

For condition (E1), it is useful to recall that for high-dimensional parametric models with d parameters, the square root of the offset complexity scales as $\sqrt{d/n}$, so the condition above requires the dimension d is $o(\sqrt{n})$; see Liang et al. (2015) and Bach (2024) for further discussion and bounds for other classes of nonparametric learners. Condition (E2) depends on the specification of the model $\mathscr{F}$ as well as the quality of the fine-tuned embeddings $\hat{\phi}$. It also tells us that fine-tuning should be done with the goal of predicting the labels Y_{it}; the better we do this, the more plausible condition (E2) becomes.

Proposition 2. *Work with the setup in this section. Assume that the data* $W_{it} = (Q_{it}, P_{it}, X_{it}^{in})$*, are identically distributed across i for each* $t = 1, .., T$*, where* T *is fixed. Assume conditions (E1) and (E2) hold and that* Y_{it} *and* $\mathscr{F}$ *are bounded. Then the empirical risk minimizer* $\hat{\gamma}_t^Y(S_{it}(\hat{\phi}))$ *learns* $\mathrm{E}[Y_{it} \mid S_{it}]$ *at the rate* $o_p(n^{-1/4})$*:* $\max_t \sqrt{\mathrm{E}[\{\mathrm{E}[Y_{it} \mid S_{it}] - \hat{\gamma}(S_{it}(\hat{\phi}))\}^2 \mid \hat{\phi}]} = o_p(n^{-1/2})$.

Proof. First note that for any $\gamma \in \mathscr{F}$, the random variable $\hat{Z}_{it}(\gamma) \coloneqq \gamma(S_{it}(\hat{\phi}))$ belongs to the subspace $E \subset L^2$ consisting of functions that are measurable w.r.t. the pair $(S_{it}, S_{it}(\hat{\phi}))$. Thus we have

$$\mathrm{E}\{Y_{it} - \hat{Z}_{it}(\gamma)\}^2 = \mathrm{E}\{Y_{it} - \mathrm{E}[Y_{it}|S_{it}, S_{it}(\hat{\phi})]\}^2 + \mathrm{E}\{\mathrm{E}[Y_{it}|S_{it}, S_{it}(\hat{\phi})] - \hat{Z}_{it}(\gamma)\}^2 \tag{8}$$

by the Pythagorean theorem, since $\mathrm{E}[Y_{it}|S_{it}, S_{it}(\hat{\phi})]$ is the orthogonal projection of Y_{it} onto E. It follows that the left-hand side depends on γ only through: $\mathrm{E}\{\mathrm{E}[Y_{it}|S_{it}, S_{it}(\hat{\phi})] - \hat{Z}_{it}(\gamma)\}^2 = \mathrm{E}\{\mathrm{E}[Y_{it}|S_{it}] - \hat{Z}_{it}(\gamma)\}^2$, where we use the fact that S_{it} is derived from the ideal embeddings.

Now, let γ_t^* in the closure of $\mathscr{F}$ satisfy $\mathrm{E}\{Y_{it} - \hat{Z}_{it}(\gamma_t^*)\}^2 = \inf_{\gamma \in \mathscr{F}} \mathrm{E}\{Y_{it} - \hat{Z}_{it}(\gamma)\}^2$. By (8) above, the same γ_t^* also satisfies $\mathrm{E}\{\mathrm{E}[Y_{it}|S_{it}] - \hat{Z}_{it}(\gamma_t^*)\}^2 = \inf_{\gamma \in \mathscr{F}} \mathrm{E}\{\mathrm{E}[Y_{it}|S_{it}] - \hat{Z}_{it}(\gamma)\}^2$, which is $o_p(n^{-1/2})$ by (E2).

Next, by Theorem 3 of Liang et al. (2015) combined with Markov's inequality, $\mathrm{E}\{Y_{it} - \hat{Z}_{it}(\hat{\gamma}_t^Y)\}^2 \leq \mathrm{E}\{Y_{it} - \hat{Z}_{it}(\gamma_t^*)\}^2 + o_p(n^{-1/2})$. Decomposing both expectations using (8), we recover

$$\mathrm{E}\{\mathrm{E}[Y_{it}|S_{it}] - \hat{Z}_{it}(\hat{\gamma}_t^Y)\}^2 \leq \mathrm{E}\{\mathrm{E}[Y_{it}|S_{it}] - \hat{Z}_{it}(\gamma_t^*)\}^2 + o_p(n^{-1/2}) = o_p(n^{-1/2})$$

where the last step follows by the previous paragraph. To conclude, we take a union bound over finitely many periods t to deduce that $\max_t \mathrm{E}\{\mathrm{E}[Y_{it}|S_{it}] - \hat{Z}_{it}(\hat{\gamma}_t^Y)\}^2 = o_p(n^{-1/2})$, as needed. □

We describe the DML slope estimator next. Let $(I_\ell)_\ell^L$ be the partition of $[n] = \{1, ..., n\}$ into L folds of approximately equal size. In step 1, for each ℓ: obtain $\hat{\gamma}_{t,\ell}^Y(S_{it}(\hat{\phi}))$, the empirical risk minimizer over observation indices $A = [n] \setminus I_\ell$, for $Y = Q$ and $Y = P$; then obtain the residuals

$\hat{P}_{it}^{\perp} = P_{it} - \hat{\gamma}_{t,\ell}^{P}(S_{it}(\hat{\phi}))$ and $\hat{Q}_{it}^{\perp} = Q_{it} - \hat{\gamma}_{t,\ell}^{Q}(S_{it}(\hat{\phi}))$ for $i \in I_\ell$. Then, obtain the slope estimator as:

$$\hat{\delta}_t = (\sum_{i=1}^{n} (\hat{P}_{it}^{\perp})^2)^{-1} \sum_{i=1}^{n} \hat{P}_{it}^{\perp} \hat{Q}_{it}^{\perp}.$$

Then, appealing to the theoretical results in Chernozhukov et al. (2018a) for partially linear models, we obtain $\sqrt{n}$-consistency and asymptotic normality for δ_t.

Corollary 1. *Work with conditions of the previous proposition. In addition, assume that* $\mathrm{E}P_{it}^{\perp 2}$ *is bounded away from zero. Then* $\sqrt{n}(\hat{\delta}_t - \delta_t) = (\mathrm{E}P_{it}^{\perp 2})^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} P_{it}^{\perp} e_{it} + o_p(1) \to_d N(0, V)$, *where* $V = \mathrm{E}[P_{it}^{\perp 2}]^{-2} \mathrm{E}[P_{it}^{\perp 2} e_{it}^2]$.

In summary, with a sufficiently large sample size for fine-tuning and highly informative embeddings, it is plausible that the projection function $\mathrm{E}[Y_{it} \mid S_{it}]$ can be approximated well enough to support standard asymptotic inference.

References

S. Athey and G. W. Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11(1):685–725, 2019.

F. Bach. *Learning theory from first principles*. MIT press, 2024.

P. Bach, V. Chernozhukov, M. S. Kurz, and M. Spindler. DoubleML: An object-oriented implementation of double machine learning in Python. *The Journal of Machine Learning Research*, 23(1):2469–2474, 2022.

P. Bach, M. S. Kurz, V. Chernozhukov, M. Spindler, and S. Klaassen. Doubleml: An object-oriented implementation of double machine learning in R. *Journal of Statistical Software*, 108:1–56, 2024.

P. Bajari, Z. Cen, V. Chernozhukov, et al. Hedonic prices and quality adjusted price indices powered by AI. *arXiv preprint, arXiv:2305.00044*, 2023.

H. Bao, L. Dong, and F. Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

A. Belloni, V. Chernozhukov, and C. Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014.

T. Brown, B. Mann, N. Ryder, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

V. Chernozhukov, D. Chetverikov, M. Demirer, et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018a. ISSN 1368-4221.

V. Chernozhukov, I. Fernández-Val, and Y. Luo. The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica*, 86(6):1911–1938, 2018b.

V. Chernozhukov, C. Cinelli, W. Newey, A. Sharma, and V. Syrgkanis. Long story short: Omitted variable bias in causal machine learning. *arXiv preprint, arXiv:2112.13398*, 2021.

J. Chevalier and A. Goolsbee. Measuring prices and price competition online: Amazon. com and barnesandnoble. com. *Quantitative Marketing and Economics*, 1:203–222, 2003.

G. Compiani, I. Morozov, and S. Seiler. Demand estimation with text and image data. *SSRN Working Paper*, 2023.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

R. Frisch and F. V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1 (4):387–401, 1933.

Z. Griliches. Hedonic price indexes for automobiles: an econometric analysis of quality change. In *Price indexes and quality change: Studies in new methods of measurement*, pages 55–87. Harvard University Press, 1971.

R. Z. Hasminskii and I. A. Ibragimov. On the nonparametric estimation of functionals. In *Proceedings of the 2nd Prague Symposium on Asymptotic Statistics*, pages 41–51, 1978.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

S. He and B. Hollenbeck. Sales and rank on Amazon. com. *SSRN Working Paper*, 2020.

W. B. Johnson. Extensions of Lipshitz mapping into Hilbert space. In *Conference in modern analysis and probability, 1984*, pages 189–206, 1984.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

S. Klaassen, J. Teichert-Kluge, P. Bach, V. Chernozhukov, M. Spindler, and S. Vijaykumar. DoubleMLDeep: Estimation of causal effects with multimodal data. *arXiv preprint, arXiv:2402.01785*, 2024.

K. H. Lee and L. Musolff. Entry into two-sided markets shaped by platform-guided search. *Preprint*, 2022. URL `https://conference.nber.org/confer/2022/DTs22/judy1.pdf`.

B. Y. Levit. On efficiency of a class of non-parametric estimates. *Teoriya Veroyatnostei i ee Primeneniya*, 20(4):738–754, 1975.

T. Liang, A. Rakhlin, and K. Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1260–1285. PMLR, 2015.

Y. Liu, M. Ott, N. Goyal, et al. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint, arXiv:1907.11692*, 2019.

D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados. TimeLMs: Diachronic language models from Twitter. *arXiv preprint, arXiv:2202.03829*, 2022.

M. C. Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR*, 2013.

F. C. Mills. On the use of index numbers of prices in the study of economic changes. *Journal of the American Statistical Association*, 26(173A):116–119, 1931.

F. C. Mills. Industrial productivity and prices. *Journal of the American Statistical Association*, 32(198): 247–262, 1937a.

F. C. Mills. Movements of mail-order prices. *Journal of the American Statistical Association*, 32(197): 131–132, 1937b.

S. Mullainathan and J. Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2): 299–319, 2021.

U. K. Office of National Statistics. Using statistical distributions to estimate weights for web-scraped price quotes in consumer price statistics. Technical report, 2020.

A. Pakes. A reconsideration of hedonic price indexes with an application to PC's. *American Economic Review*, 93(5):1578–1596, 2003.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

J. Pfanzagl and W. Wefelmeyer. Contributions to a general asymptotic statistical theory. *Statistics & Risk Modeling*, 3(3-4):379–388, 1985.

I. Reimers and J. Waldfogel. Digitization and pre-purchase information: The causal and welfare impacts of reviews and crowd ratings. *American Economic Review*, 111(6):1944–71, June 2021.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

D. B. Rubin. Bayesian inference for causality: The importance of randomization. In *The Proceedings of the social statistics section of the American Statistical Association*, volume 233, page 239. American Statistical Association Alexandria, VA, 1975.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

H. Schultz. The standard error of the coefficient of elasticity of demand. *Journal of the American Statistical Association*, 28(181):64–69, 1933.

S. Seabold and J. Perktold. Statsmodels: econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*, 7(1), 2010.

V. Semenova, M. Goldman, V. Chernozhukov, and M. Taddy. Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics*, 14(2):471–510, 2023.

G. Somepalli, A. Schwarzschild, M. Goldblum, et al. SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. In *NeurIPS 2022 First Table Representation Workshop*, 2022.

G. J. Stigler. The limitations of statistical demand curves. *Journal of the American Statistical Association*, 34(207):469–481, 1939.

H. Touvron, T. Lavril, G. Izacard, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint, arXiv:2302.13971*, 2023.

H. R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28, 2014.

A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

H. Working. Statistical laws of family expenditure. *Journal of the American Statistical Association*, 38 (221):43–56, 1943.

P. G. Wright. Review: Statistical laws of demand and supply by Henry Schultz. *Journal of the American Statistical Association*, 24(166):207–215, Jun. 1929.

P. G. Wright. Effects of a duty on price and output with special reference to butter and flaxseed. *The Econometrics Journal*, page utae018, 12 2024.

J. R. Zaurin and P. Mulinka. pytorch-widedeep: A flexible package for multimodal deep learning. *Journal of Open Source Software*, 8(86):5027, 2023.

Online Appendix

to

Adventures in Demand Analysis Using AI

B Workflow

The workflow is based on the dataset $(\text{Data}_{it})_{i\in I, t\in\{0,\dots T\}}$, described in Section 2.1, where I denotes the set of all collected product IDs ("ASIN") and $\{0,\dots T\}$ the observed time periods.

1. **Dataset Split:** Divide the dataset $(\text{Data}_{it})_{i\in I, t\in\{0,\dots,T\}}$ into two disjoint subsets based on product IDs: $I = I_1 \cup I_2$.

2. **Embedding Fine-Tuning:** Fine-tune the high-dimensional product embeddings $\tilde{E}_i$ on I_1 by training the neural network described in Section C using the following loss function:

$$L = \|Q_{it} - \hat{l}(X_i^{in})\|_{pred,2} \cdot \|P_{it} - \hat{m}(X_i^{in})\|_{pred,2},$$

   where $\|\cdot\|_{pred,2}$ is the empirical root mean square error (empirical prediction norm).

3. **Embedding Computation:** Compute high-dimensional product embeddings $\tilde{E}_i \in \mathbb{R}^{1888}$ for all product IDs $i \in I$.

4. **Feature Computation:** Following Algorithm 1, compute 256-dimensional embeddings E_i as well as PCA-based and similarity features $PC_{i,k}$ and $CS_{i,k}$ for $k = 1,\dots,5$. Perform this computation for all observations (i,t) with $i \in I$ and $t \in \{0,\dots,T\}$.

5. **Partialling-Out:** For all observations (i,t) where $i \in I_2$ and $t \in \{1,\dots,T\}$, partial out the effects of $(Q_{i,t-1}, P_{i,t-1}, X_i^{in}, X_{it}^{o})$ to obtain $\hat{Q}_{it}^{\perp}$ and $\hat{P}_{it}^{\perp}$:

(a) Use a **linear specification** (Algorithm 2) controlling for either $V_i(X_i^{in}) = CS_i$ or $V_i(X_i^{in}) = E_i$.

(b) Use a **partially linear specification** (Algorithm 3) controlling for either $V_i(X_i^{in}) = CS_i$ or $V_i(X_i^{in}) = E_i$.

6. **Homogeneous Effect:** Estimate the homogeneous effect $\hat{\alpha}$ by performing a linear regression: $\hat{Q}_{it}^{\perp} \sim \hat{P}_{it}^{\perp}$, clustering the standard errors at the product ID level ("ASIN").

7. **Heterogeneous Effect:** To estimate the heterogeneous effect $\hat{\alpha}(X_{it}^{e})$, perform a linear regression:

$$\hat{Q}_{it}^{\perp} \sim \hat{P}_{it}^{\perp} \cdot \left(1 + Q_{i,t-1} + P_{i,t-1} + CS_{i1} + CS_{i2} + CS_{i3} + CS_{i4} + CS_{i5}\right),$$

clustering the standard errors at the product ID level ("ASIN").

The PCA algorithm and cluster similarity analysis, including k-means, are implemented using the `scikit-learn` package (Pedregosa et al., 2011). Linear regression models are performed using the `statsmodels` package (Seabold and Perktold, 2010). For the estimation of partially linear models, the `DoubleML` package (Bach et al., 2022, 2024) is used, with boosting algorithms implemented via the `Lightgbm` package (Ke et al., 2017). To simplify computations in high-dimensional settings, the embeddings are projected to a lower-dimensional space using Gaussian Random Projection, which preserves distances as described in (Johnson, 1984).

Algorithm 1: Generating PCA and Similarity Features

Input: Embeddings $E_i \in \mathbb{R}^d$ for $i \in I$.

Output: Projected and Normalized Embeddings $X_i^e \in \mathbb{R}^{256}$, Cluster Centroids CS_i, and Principal Components PC_i.

1 Project embeddings E_i onto a lower-dimensional space $\bar{E}_i \leftarrow E_i^T G$, where $G \in \mathbb{R}^{d\times 256}$ is a random matrix with i.i.d. N(0,1) entries;

2 Center and normalize embeddings: $X_i^e \leftarrow \frac{\bar{E}_i - \frac{1}{n}\sum_{i\in I}\bar{E}_i}{\|\bar{E}_i - \frac{1}{n}\sum_{i\in I}\bar{E}_i\|_2}$;

4 **Step 1: PCA Features**;

- Run a PCA algorithm on $(X_i^e)_{i\in I}$;
- Project each X_i onto the first k-th principal component $PC_{ik} := PC_k(X_i^e)$ for $k = 1,\ldots,K$;

5 **Step 2: Similarity Features**;

- Use k-means clustering with euclidean distance on $(X_i^e)_{i\in I}$ to compute K centroids c_k;
- Compute cosine similarities $CS_{ik} := CS_k(X_i^e) = \frac{c_k^T X_i^e}{\|c_k\|_2\|X_i^e\|_2}$ for $k = 1,\ldots,K$;

6 Return X_i^e, $CS_i = (CS_{i1},\ldots,CS_{iK})$ and $PC_i = (PC_{i1},\ldots,PC_{iK})$;

Algorithm 2: Partialling-out: Linear Specification

Input: Data $(Q_{it}, P_{it}, Q_{i,t-1}, P_{i,t-1}, V_i, X_{it}^o)$ for $i \in I_2$, $t \in \{1,\ldots T\}$

Output: Partialling out values $\hat{Q}_{it}^{\perp}$, $\hat{P}_{it}^{\perp}$

1 Run ordinary least squares

$$Q_{it} \sim Q_{i,t-1} + P_{i,t-1} + V_i + X_{it}^o, \qquad P_{it} \sim Q_{i,t-1} + P_{i,t-1} + V_i + X_{it}^o,$$

and keep predicted values $\hat{Q}_{it}$ and $\hat{P}_{it}$.

2 Output the residuals $\hat{Q}_{it}^{\perp} := Q_{it} - \hat{Q}_{it}$, and $\hat{P}_{it}^{\perp} := P_{it} - \hat{P}_{it}$.

Algorithm 3: Partialling-out: Partially Linear Specification

Input: Data $(Q_{it}, P_{it}, Q_{i,t-1}, P_{i,t-1}, V_i, X_{it}^o)$ for $i \in I_2$, $t \in \{1, \dots T\}$

Output: Partialling out values $\hat{Q}_{it}^{\perp}$, $\hat{P}_{it}^{\perp}$

1 Employ 5-fold cross-fitting to estimate corresponding non-linear regression functions

$$l_0(Q_{i,t-1}, P_{i,t-1}, V_i, X_{it}^o) := \mathrm{E}[Q_{it} | Q_{i,t-1}, P_{i,t-1}, V_i, X_{it}^o]$$

$$m_0(Q_{i,t-1}, P_{i,t-1}, V_i, X_{it}^o) := \mathrm{E}[P_{it} | Q_{i,t-1}, P_{i,t-1}, V_i, X_{it}^o]$$

via boosted trees.

2 Compute and report the residuals

$$\hat{Q}_{it}^{\perp} := Q_{it} - \hat{l}_0(Q_{i,t-1}, P_{i,t-1}, V_i, X_{it}^o)$$

$$\hat{P}_{it}^{\perp} := P_{it} - \hat{m}_0(Q_{i,t-1}, P_{i,t-1}, V_i, X_{it}^o)$$

C Neural Network Architecture

To model demand based on multimodal data inputs, we develop and implement an architecture presented in Figure 9. As our interest is not only in predicting the quantity demanded Q_{it}, but also on valid statistical inference on elasticity parameters, our architecture has been adapted to the double/debiased machine learning framework (Chernozhukov et al., 2018a). The later is based on an orthogonal moment condition to account for typical machine-learning induced biased and, hence, requires to additionally perform predictions on the price variable P_{it}, see Klaassen et al. (2024) for more details. Each of these transformer blocks will output a dense embedding, which are later being concatenated to the multimodal embedding E_i.

The following model components were used for implementation: The SAINT model (Somepalli et al., 2022) implemented in the `pytorch-widedeep` package (Zaurin and Mulinka, 2023) has been used as the tabular encoder, RoBERTa model being pretrained on a Twitter Dataset (Loureiro

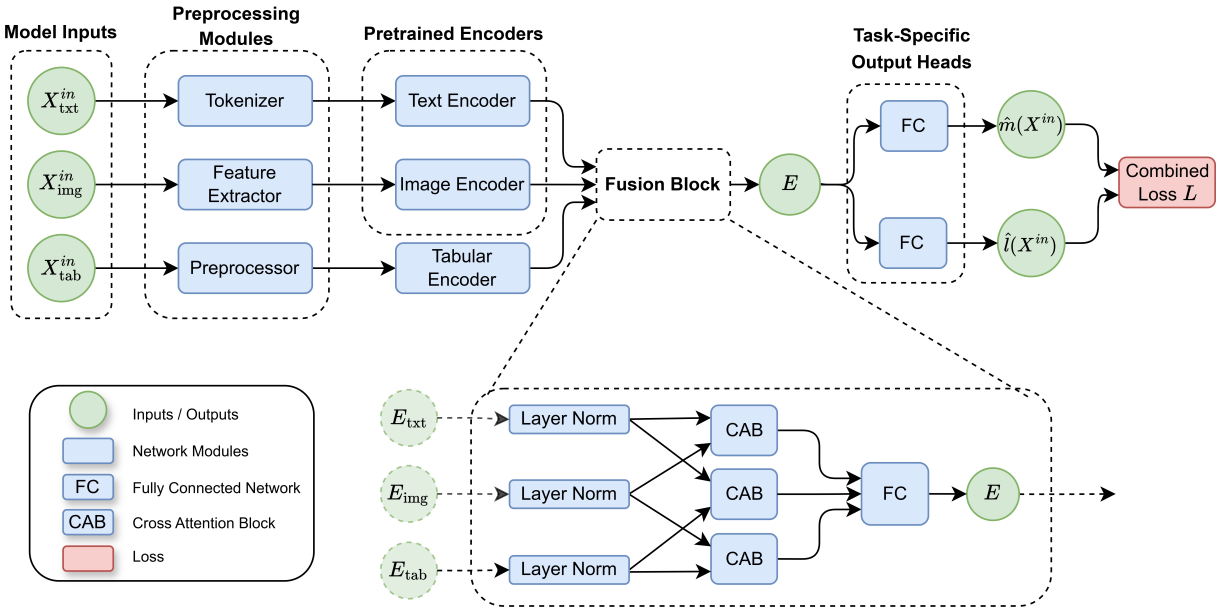

Figure 9: Proposed model architecture

et al., 2022) as the text encoder, and the BEiT model (Bao et al., 2022) as the image encoder.