Using Contextual Embeddings to Predict the Effectiveness of Novel Heterogeneous Treatments

Paul B. Ellickson, Wreetabrata Kar, James C. Reeder, III, and Guang Zeng*

May 28, 2024

Abstract

Our study demonstrates the power of contextual embeddings for predicting the performance of novel heterogeneous treatments. Our proposed framework leverages four key benefits of machine learning: prediction, enhanced causal estimation, estimation of heterogeneous effects, and generative capability. To test our framework, we exploit a targeted marketing setting in which 34 email promotions were sent to 1.3 million customers over a 45-day period. Using these emails as treatments, we start by estimating the doubly robust scores of customer-level purchase amounts to serve as our target variable. We incorporate customer-level demographics and contextual embeddings, which capture the context of the latent states, to estimate the response function of these emails. Using a series of leave-one-out exercises, we show how our approach can accurately extrapolate the average performance, heterogeneous performance, and recommended targeting policies of novel promotions. We find that our framework recovers 78.6% of the variation of the aggregate treatment effects, an average of 65.36% of the variation in the heterogeneous treatment effects of each novel treatment, and matches 82% of policy recommendations made using the true signals. We conclude our study with an example of leveraging Generative AI to create novel treatments and then evaluate their performance with our framework.

Keywords: Contextual Embeddings, Digital Marketing, Double Machine Learning, Multi-Treatment Heterogeneity, Large Language Models, ChatGPT.

*Ellickson: Simon School of Business, University of Rochester, paul.ellickson@simon.rochester.edu. Kar: Krannert School of Management, Purdue University, wkar@purdue.edu. Reeder: School of Business, University of Kansas, jcreederiii@ku.edu. Zeng: Simon School of Business, University of Rochester, guang.zeng@simon.rochester.edu.

1 Introduction

"So how is it, then, that something like ChatGPT can get as far as it does with language? The basic answer, I think, is that language is at a fundamental level somehow simpler than it seems. And this means that ChatGPT is successfully able to "capture the essence" of human language and the thinking behind it."

- Stephen Wolfram, 2023.[1]

Machine learning (ML) algorithms are the leading empirical approach to prediction and pattern recognition across many domains. In addition, causal ML methods now play an increasingly important role in providing pathways to credible causal inference, under the assumption of selection on observables, by allowing for data-driven model selection over both confounders and effect modifiers (Chernozhukov et al., 2024). In the latter case, they have also shown tremendous promise in detecting and quantifying heterogeneity in treatment effects. Most recently, machine learning techniques associated with deep neural nets and transformer architectures have been used to capture, in a sparse yet complete way, the key meaning and content in prose, irrespective of length or complexity. The contextual information inferred from language can then be used in a host of down-stream tasks, collectively known as generative Artificial Intelligence (GenAI).

In this paper, we integrate these four key strengths of machine learning — prediction, causal inference, estimation of heterogeneous effects, and generative capabilities — to predict the performance of new marketing creatives (treatments). Our application, which builds upon and extends Ellickson et al. (2023), compares the actual and predicted performance of targeted email promotions designed to elicit incremental purchases. The contextual embeddings that drive modern large language models enable us to obtain counterfactual predictions. Importantly, these embeddings are situated on a common linguistic manifold, allowing for extrapolation. Combined with doubly robust ML techniques, we estimate unbiased signals

[1] `https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/`

of counterfactual performance based on the past performance of several comparable creatives (i.e., email promotions in our context), which is key to predicting the success of new creatives. Moving beyond average prediction, we further leverage recent advances in the use of sorted group average treatment effects to identify treatment effect heterogeneity within each creative, an important input to designing targeting heuristics. Classification analysis then ties the heterogeneity to observable demographic variables.[2] Finally, we employ policy-learning approaches to leverage the connection to demographics in the service of identifying better targeting policies. At each step, out-of-sample performance is demonstrated via our collection of observed creatives via a suite of leave-one-out exercises.[3]

We begin by proposing a framework for conducting both inference and prediction on latent, contextual treatments characterized by heterogeneous textual information. Building upon Egami et al. (2022), we introduce four key assumptions required for causal prediction, along with a strategy for constructing predictions, identifying heterogeneity, and assigning treatments. We emphasize the central role of doubly robust scores of relative outcome performance in serving as the focal building block of three downstream post-processing tasks.[4] First, their predicted analogs can be averaged to characterize the average performance of new treatments. Second, their sorted values can be grouped to identify heterogeneity in the response to these treatments and further connected to the demographics that index it. Third, the doubly robust scores can facilitate policy-learning-based approaches to targeting new treatments, bypassing the need for new experiments.

After developing our causal prediction framework, we then turn to the task of dimension reduction, highlighting the key role that embeddings play in capturing the content of complex treatments in marketing. We discuss the benefits and trade-offs of alternative encoders and

[2]In our context, demographic variables relate to both the demographic characteristics of customers and the recency-frequency-monetary variables that capture customers' interactions with the firm.

[3]We choose this approach to ensure that our estimation of the response surface does not include any direct information about the focal treatment (or creative), thus offering a rigorous benchmark for comparing performance in the development of novel treatments.

[4]Doubly robust scores, the individual-level estimates coming from a double machine learning (DML) estimator, are also called doubly robust signals or individual-treatment effects (ITEs).

the importance of sample-splitting to avoid over-fitting. The use of pre-trained encoders can do so automatically. Turning to the task of prediction, namely the mapping from embeddings to doubly robust outcome signals, we formalize the requirements for predictive accuracy and discuss procedures for additional dimension reduction in settings with low power. We target a predictive response surface whose features can then be exploited for downstream targeting.

We demonstrate the capability and flexibility of our framework on a set of 34 email promotions delivered to a target population of 1.3 million customers. Using the estimation procedure proposed in Ellickson et al. (2023), we first calculate doubly robust scores for 33 emails. Each score denotes the customer-specific increase in purchase amount that each email generates compared to a fixed baseline email. We then employ a sorted group average treatment effect (GATES) approach (Chernozhukov et al., 2017, 2018) to document significant heterogeneity in the response to these individual emails.[5] In our out-of-sample tests of predictive performance for emails that were not used to train the model, we rely on these doubly robust signals and GATES constructs to capture the ground truth. Through a sequence of out-of-sample exercises (leave-one-email-out), we showcase our framework's capability to predict not only the average performance of a new email but also to discern heterogeneity in its underlying causal effect, which strongly correlates with the ground truth provided by the scores. For example, examining the average performance (ATE) of novel treatments, we find that we are only unable to accurately predict the performance of 2 out of 33 emails and our approach captures 78.6% of the variation in estimated ATEs. Further, exploring the heterogenous effects of the novel treatments using the GATEs, we find that our approach captures an average 65.36% of the variation in the heterogenous effects of the novel treatments. This heterogeneity, which is further tied to demographic variables capturing the engagement and revenue potential of the target customer population, is the key requirement for successful targeted marketing. We then demonstrate how the same predictive outputs of our framework can be used for policy learning to make such targeting operational. Once

[5]Though not explicitly referred to as GATES in Ascarza (2018), the use of sorted treatment effects in this marketing application is crucial for determining to whom treatment should be administered.

again, out-of-sample evaluation demonstrates that working with the counterfactual predictions retains much of the signal of the underlying ground truth.

In our last empirical exercise, we circle back to the potential of generative AI by evaluating ChatGPT's ability to both ideate new creatives and predict their performance. Similar to recent studies, we find that ChatGPT excels at creation but struggles at evaluation (Girotra et al., 2023; da Costa et al., 2023), which may reflect its tendency to hallucinate responses (Ji et al., 2023). Recent research suggests that fine tuning a large language model with additional data may even exacerbate the hallucination problem, which may make the new model even more unreliable (Gekhman et al., 2024). In our application, ChatGPT is able to suggest new email creatives that have varying levels of performance; however, it fails to rank the relative performance of these creatives in an informative manner. Thus, we advocate for ChatGPT to generate a panel of new creatives and then use our approach, based on firm-level data, in describing the response surface of treatment characteristics to validate the performance of the new creatives.

Our paper resides in the nexus of two streams of literature: leveraging machine learning for better policy outcomes and using large language models/contextual embeddings in business contexts. The use of machine learning for marketing applications has exploded in recent times. While we do not provide an exhaustive list of such studies, we highlight a few examples where machine learning has been utilized in this capacity. Ascarza (2018) uses machine learning to assess the retention rate of individuals in order to best target the use of promotions. Yoganarasimhan et al. (2020) leverages an experiment to find the optimal length of a free trial for a software product. Dubé and Misra (2023) focuses on finding the profit maximizing price through an experiment. Further studies have looked at development of email creatives, most closely related to our empirical setting, using machine learning (Balakrishnan and Parekh, 2014; Jaidka et al., 2018; Nguyen et al., 0).

The prior studies have either focused on low-dimensional treatment objects (mostly binary), or codification of treatment characteristics into an easier to manipulate space. With

the creation of large language models and transformer architecture, the text associated with treatments can either be operationalized directly through embeddings or through fine-tuning of existing large language models. For the former case, Bajari et al. (2023) use text embeddings to control for product information as a means to better assess the impact of price. Dew (2023) uses PCA-reduced image embeddings for the purposes of developing new visual images. Reisenbichler et al. (2022) is one of the first papers to use business outcomes in fine-tuning a large language model with the goal of enhancing SEO performance. Most closely resembling our study is Angelopoulos et al. (2024), which proposes fine-tuning a large language model to generate new email creatives and evaluating their performance across a large collection of RCTs. In contrast, our paper focuses on combining the effectiveness of contextual embeddings to capture high-dimensional information with customer demographics and treatment outcomes to extrapolate the individual-level performance of novel heterogeneous treatments. Consequently, we can not only predict the average performance of new treatments or creatives, as demonstrated in Angelopoulos et al. (2024), but also assess heterogeneous responses and provide targeting guidelines.

The rest of the paper is organized as follows. In Section 2 we propose a causal inference framework with latent treatments, which guides researchers to choose machine learning algorithms for generating latent treatments, and conduct predictive tasks for effects of new treatments. We also highlight the usefulness of the contextual embeddings for our application. Section 3 describes the institutional context and construction of the dataset. Section 4 contains our results across three stages of analysis: the aggregate, average response to novel treatments, the heterogeneous response to novel treatments, and the comparability in policy performance of novel treatments. Further, in Section 5 we present an empirical exercise that connects ChatGPT ideation with our framework to vet the performance of novel treatments. Section 6 concludes.

2 Causal Inference with Latent Contextual Treatments

A key challenge in using unstructured data, regardless of form, is operationalizing it into a useful construct for analysis. The easiest and most common method of encoding unstructured text is a categorical encoder: the simple assignment of each treatment to a unique integer identifier. The benefit of this approach lies in allowing the researcher to work under a standard multivalued treatment framework that permits *causal inference* tasks. However, categorical encoders do not allow researchers to predict (extrapolate) a new counterfactual scenario that was not observed in the data (i.e., novel treatments).[6] The ability to extrapolate to new treatments requires the encoding of information contained in the unstructured treatments into a latent space that allows for a universal representation of the treatment's characteristics, both observed and hypothetical. Recent progress in machine learning algorithms now allows researchers to encode unstructured data at scale. As a result, researchers have started to conduct causal inference on encoded treatments.

Egami et al. (2022) and Fong and Grimmer (2023) provide useful guidance on text as treatments, framing the problem as one of causal inference on latent treatments. Our approach builds on these concepts, further formalizing and extending them within the potential outcomes causal framework to counterfactual prediction tasks. The core idea of our approach is to convert the observed unstructured treatments to structured contextual treatments and define the treatment response surface as a function of these contextual factors. We then predict the effect of a new unstructured treatment through its structured contextual correspondence.

In this section, we begin by formalizing the task of using latent treatments in a causal framework. Next, we link these latent treatments to causal inference by employing doubly robust signals of their causal effects. Last, we highlight how latent treatments and estimated doubly robust scores can be used to predict the effectiveness of novel heterogeneous

[6]This problem is similar to adding a new product to a standard (characteristics-based) discrete choice demand system. If that new product includes a feature not observed in the current product set, prediction of its counterfactual performance is intractable.

treatments in contextual space.

2.1 Causal Framework with Latent Treatments

We now introduce a framework that formalizes this practice. We also discuss the requirements for choosing the associated encoding function, and provide concrete guidance for doing so. In presenting the framework, we use textual information as our main example of unstructured treatments to illustrate the main concepts. However, these concepts apply to any unstructured information that can now be encoded and projected onto a common, low-dimensional manifold (e.g., images, video, and audio).

We start from the premise that many unstructured data, such as texts, reflect an underlying structured contextual *meaning*, which we refer to as its context. To fix ideas, consider the set of all possible (English) text documents that currently exist, denoted as Δ, and the set of all possible contextual meanings expressed by these texts, denoted as Γ. A human language system defines a mapping from texts to contexts, $H : \Delta \to \Gamma$. Notably, the elements in Δ, texts, are generally ultra-high-dimensional objects, while the elements in Γ, contexts, are, ideally, dimension-reduced objects capturing the same base meaning. The scope for dimension reduction arises from the apparent "structural sparsity" of language, which was alluded to in the opening quote of this paper. The reduced dimensions represent linguistically irrelevant information that is second order for capturing the underlying meaning.

Typically, when researchers are conducting causal inference with unstructured treatments, the possible set of treatments is restricted to their focal domain. For instance, in our email application, the set of treatments are all possible email subject lines sent to customers of an online women's retailer. We can then define a set of all possible treatments under consideration as $T \subset \Delta$, such that all attention can be focused on the mapping H on T, which we denote as H_T in what follows.[7] Using the above notation, we can proceed to developing

[7] T should be interpreted as being the set of both the treatments observed in the sample, and the treat-

our approach for conducting causal inference on unstructured treatments under the potential outcome framework. We start with two assumptions that characterize treatments and their associated outcomes.

Assumption 1. *For all individuals ω and any j, $k \in T$, if $j = k$, then $H_T(j) = H_T(k)$.*

Assumption 2. *For all individuals ω and all $k \in T$, potential outcomes $Y(k, \omega) = Y(H_T(k), \omega)$.*

Assumption 1 requires that, in the focal domain (T), the mapping (H_T) from observed unstructured treatments to the corresponding contexts is a one/many-to-one mapping for all individuals ω. Substantively, this requires that all individuals agree on the underlying meaning of a given textual input (though how they *react* to that text can differ). For example, this assumption rules out ambiguity in the meaning of email subject lines considered in Ellickson et al. (2023), which in their context, appears mild. Note that this is a component of the stable unit treatment value assumption (SUTVA), which asserts that there are no different versions of each treatment (Imbens and Rubin, 2015). If an observed unstructured treatment can be mapped to different contexts, then there are hidden variations of this treatment, which lead to SUTVA violations. Assumption 2 states that the observed unstructured treatments affect outcomes only through contextual information. Fong and Grimmer (2023) impose a similar assumption on their "codebook function", which requires either the absence of unmeasured latent treatments (confounders) or that any such unmeasured constructs are independent of either the measured treatments or their associated outcomes. Intuitively, this essentially requires that contextual meaning is a sufficient statistic for the complete text in question.

We now state two well-known assumptions that ensure the identification of the causal effects of our latent contextual treatments. To allow for counterfactual prediction over out-of-sample treatments, we further introduce an object $S \subset T$, which is the set of all treatments

ments not in sample, but of interest to the analyst.

observed in the data. Counterfactual prediction (extrapolation) is conducted for some treatment $l \in T \setminus S$, i.e., new cases that fall outside the sample.

Assumption 3 (Unconfounded Assignment)**.** *For all individuals ω and all $k \in S \subset T$, $Y(k, \omega) \perp\!\!\!\perp D(\omega) \mid W_\omega$, where $D(\cdot)$ is a treatment assignment function, whose range is S, and W_ω is a set of (possibly high-dimensional) observable variables attached to individual ω.*

Assumption 4 (Sufficient Overlap)**.** *For all $k \in S$, we have $0 < Pr(D = k|W) < 1$.*

Assumption 3 states that treatment assignment is orthogonal to potential outcomes, conditional on the observed demographic features W of the target population. Stated differently, if we consider individuals with the same values of observed features, namely $W = w$, then treatment variation among these individuals, $D|W = w$, is generated as if by a conditionally randomized control trial. In classical targeted marketing settings, this assumption is guaranteed by design (even if the assignment rules are, in fact, deterministic), so long as the researcher observes the same features used by firms in their targeting decisions. Assumption 4 further requires that there is sufficient random variation in D at each value w in the support of W (all treatments have some probability of being delivered). Substantively, this requires some form of randomization, either explicit or by chance. In the case of an RCT, overlap is also guaranteed by design. More generally, it often requires some form of residual randomization (after controlling for W) that was not explicitly built in to the intended assignment mechanism (e.g., due to targeting budget constraints). Under these conditions, we can then construct a contrast between any treatments j and k in S for all w in the support of W.

Assumptions 3 and 4 provide the identification conditions required for researchers to conduct causal inference within the *observed* treatment set S. In contrast, Assumptions 1 and 2 further allow for causal inference on contextual treatments in the same manner in which one estimates the causal effects of observed multivalued treatments in S under Assumptions

3 and 4. To be more precise, when researchers are predicting counterfactual outcomes (or treatment effects), the assignment rule of observed treatments (D) in sample S, act as a sampler on $H_T(S)$ in the contextual space — there is a response surface of outcomes defined on $H_T(T)$, and the assignment rule $D(\cdot)$ samples points in $H_T(S) \subset H_T(T)$. The ability to locate the position of the response surface at the sampled points in $H_T(S)$ is guaranteed by Assumption 2. Combining these sampled points on the response surface with functional form assumptions, we can interpolate/extrapolate the response surface to a new treatment value that is not in S. Based on the sampling argument, we can see that if Assumption 1 fails, e.g., there are multiple contexts corresponding to the new treatment, we cannot determine which point in the context space to use for interpolation/extrapolation, yielding an identification problem. As we have seen in the classic causal inference problems, we cannot identify causal effects when SUTVA is violated unless we impose additional assumptions or redefine the target causal estimands. For a visual depiction of the idea of counterfactual predictions (i.e., how we translate the response surface from our causal estimates into the predicted performance of novel treatments), please see Figure 1.

***** Insert Figure 1 about here *****

2.2 Causal Inference Based on Doubly Robust Scores

Satisfying Assumptions 1 – 4 allows us to conduct causal inference in the contextual treatment space. We describe that process here. Following Heckman and Vytlacil (2007), we define the following distinct treatment effects (causal estimands):

1. Individual treatment effect (ITE): $\tau_{jk}(\omega) \equiv Y(j,\omega) - Y(k,\omega) = Y(H_T(j),\omega) - Y(H_T(k),\omega)$.

2. Average treatment effect (ATE): $\tau_{jk} \equiv E_\omega[Y(j,\omega) - Y(k,\omega)] = E_\omega[Y(H_T(j),\omega) - Y(H_T(k),\omega)]$.

3. Conditional average treatment effect (CATE): $\tau_{jk}(x) \equiv E_\omega[Y(j,\omega) - Y(k,\omega)|X =$

$x] = E_\omega[Y(H_T(j), \omega) - Y(H_T(k), \omega)|X = x]$, where X is a low-dimensional vector of observable features.

Note that, in these definitions, the latter equalities are implied by Assumption 2 (potential outcomes of unstructured treatments are the same as those of corresponding latent contextual treatments). Further, note that the above treatment-effect constructs are each defined pairwise. In practice, researchers typically select a baseline treatment (e.g., treatment k) and compute the effects of each treatment compared to the common baseline k. Depending on the context, that baseline treatment can either be a true control condition or simply a convenient benchmark (e.g., a "business as usual" intervention).[8]

Although they are generally not identified econometrically, the ITE estimands serve as the fundamental building blocks for the other (aggregate) constructs. For example, the population expectation of the ITEs yields the familiar average treatment effect (ATE), while conditional expectations yield various conditional average treatment effect constructs (CATEs) that can be used as the basis for targeting.[9] Sample analogs of the ITEs are provided by doubly robust scores, which similarly serve as building blocks for the sample analogs of ATE and CATE constructs, as well as other features of the causal response surface. For example, a simple average of the scores yields a consistent estimate of ATE. Similarly, we can project the estimated ITE signals on various components or aggregates of X to obtain consistent estimates of many familiar CATE estimands. In particular, the ATE conditional on individual features of $X \subset W$, corresponds to the most familiar CATE construct.

Having discussed how the ITE signals are used to construct the various aggregates defined above, we turn now to estimating the ITEs themselves (i.e., constructing the ITE signals). To do so, we follow Robins and Rotnitzky (1995) and Chernozhukov et al. (2018) in employing doubly robust scores as our estimates of ITEs (henceforth, we use the term doubly

[8]By focusing on pairwise contrasts, it is not necessary to compute all the pairwise effects, which can reduce the computational burden.

[9]A less familiar construct is a marginal treatment effect for a treatment that itself is a bundle of features, k, where k is a d-component vector $(k_1, \ldots, k_d)$ of individual treatment components. Expectations that condition on the presence of particular components (or sets of components) yield various marginal average treatment effects (or features of the underlying response surface).

robust scores to represent ITE signals). This allows for both the use of flexible machine learning (ML) algorithms in performing estimation, and conducting valid inference over its outputs. The use of ML estimators is important here because researchers may not know the exact functional form of the assignment rule and/or the structure of the potential outcome models that characterize the response to treatments. Variable selection, in particular, may be required for each, as well as a data-driven approach to function approximation. Both introduce potential regularization biases that must then be addressed econometrically. Doubly robust scores provide a guard against this bias (see discussion below).

Following Chernozhukov et al. (2018), the doubly robust scores for individual ω take the following form:

$$\begin{aligned}\tau_{jk,dr}(\omega) = \mu(j, W_\omega) - \mu(k, W_\omega) \\ + \frac{D_j(\omega)}{e(j, W_\omega)}[Y_\omega - \mu(j, W_\omega)] - \frac{D_k(\omega)}{e(k, W_\omega)}[Y_\omega - \mu(k, W_\omega)],\end{aligned} \tag{1}$$

where $\mu(k, W)$ is the outcome model for treatment k, $D_k(\omega) = \mathbb{1}[D_\omega = k]$, $e(k, W)$ is the propensity score model for treatment k, and Y_ω is the observed outcome of individual ω. To obtain the estimated doubly robust scores, $\hat{\tau}_{jk,dr}$, we proceed in two steps. First, we randomly partition the data into approximately equal-sized folds, indexed by $i \in \{1, \cdots, I\}$. For each fold $i = 1, \cdots, I$, we compute ML estimators $\hat{\mu}_{[i]}$ and $\hat{e}_{[i]}$ of μ and e, leaving out the i-th fold of data. Second, we plug the estimated $\hat{\mu}$ and $\hat{e}$ corresponding to each observation into the doubly robust score formula, after which we can conduct corresponding downstream post-processing to obtain the estimator for the target estimand. Chernozhukov et al. (2018) establish that the doubly robust scores satisfy a Neyman orthogonality condition that ensures that they do not suffer from the regularization bias arising from ML estimation (provided sufficiently high-quality ML estimators are used to construct the nuisance functions). Further, the use of a sample-splitting procedure controls over-fitting bias. In practice, researchers may also need to trim the estimated propensity scores, $\hat{e}$, in order to reduce the disproportionate impact of extreme propensity score weights.

2.3 Predictive Inference on Contextual Space

Turning now to our task of predicting counterfactual outcomes at some out-of-sample treatment values, e.g., $l \in T \setminus S$, the estimation procedure proceeds as follows (assuming k is the baseline treatment):

1. Obtain the estimated doubly robust scores $\hat{\tau}_{jk,dr}$ for every $j \in S \setminus \{k\}$ observed in S;
2. Obtain the estimated contextual mapping $\hat{H}_T(\cdot)$;
3. Project $\hat{\tau}_{jk,dr}$ on $\hat{H}_T(\cdot)$ and W using a candidate predictive model $G_T(H_T, W)$, yielding the estimated model response surface $\hat{G}_T(\cdot, \cdot)$;
4. Predict $\hat{\tau}_{lk,dr} = \hat{G}_T(\hat{H}_T(l), W)$, as a fitted value from the estimated function recovered in step 3.

The basic workflow of the prediction exercise is as follows:

$$\text{Unstructured treatment } l \xrightarrow{\text{encoder } \hat{H}_T(\cdot)} \text{context of } l \xrightarrow{\text{predictive model } \hat{G}_T(\cdot,\cdot)} \text{(C)ATE of } l.$$

Compared with typical causal estimation tasks carried out on the observed treatments in the sample, there are several additional steps for causal prediction. First, we must estimate/recover the mapping H_T that provides a low-dimensional characterization of the context. Second, we must specify a predictive model G_T that maps from $H_T(T)$ to the doubly robust scores. Third, we must estimate G_T as a function of the recovered contexts $H_T(S)$, yielding the response surface. In the following subsections, we discuss our proposed estimation strategies for H_T and G_T.

2.3.1 Estimation of H_T

Recall that H_T is the dimension-reduced contextual space that captures the content of the treatments. There are many ways to estimate H_T (i.e. carry out the dimension reduction

step). A general requirement for constructing $\hat{H}_T$ is the use of sample-splitting: employing a different sample for dimension reduction than the one subsequently used for causal estimation. This is necessary to prevent over-fitting. Additionally, when selecting an estimator for H_T, the ideal requirement is that the estimator $\hat{H}_T$ be "close to" the true H_T, i.e., that $||H_T - \hat{H}_T||_\infty$ is small. However, this high-level condition is difficult to motivate in practice, given the dearth of formal statistical investigations into the properties of estimators for unstructured data. We thus impose a weaker requirement that can guide researchers in selecting the best $\hat{H}_T$ from a collection of different estimator options in a practical manner. This weaker requirement is, for any $j, k, l \in T$, we have

$$||H_T(j) - H_T(k)|| < ||H_T(j) - H_T(l)|| \Leftrightarrow ||\hat{H}_T(j) - \hat{H}_T(k)|| < ||\hat{H}_T(j) - \hat{H}_T(l)||,$$

where $||\cdot||$ is the Euclidean norm. The weaker condition states that the estimator $\hat{H}_T$ should preserve the order of the distance between different treatments in the contextual space, which is a requirement for the quality of an encoder.[10] For predictive purpose, this requirement is sufficient. To see why, consider a partition-based predictive technique that leverage local information to form a prediction.[11] When the order of distance between any pair of contexts is preserved by the encoder $\hat{H}_T(\cdot)$, then the set of neighbors is also preserved for any given context, which is essential to the downstream partition-based predictive algorithm's performance. In addition, practitioners often standardize $\hat{H}_T(T)$ before moving to predictive modeling — in this case, the distance-order-preserving condition is equivalent to a distance-preserving condition or a consistency condition.[12] When choosing among encoders, we should pick the one that satisfies this requirement as closely as possible.

[10]This condition is weaker than strong assumptions such as requiring that the encoder be a consistent estimator, a property for which we currently have little guidance or formal results in most cases.

[11]Many popular ML algorithms fall into this category, including random forests, (boosted) trees, and Lasso regressions with spline bases.

[12]Whether it is a distance-preserving condition or a consistency condition depends on how researchers standardize the estimated contexts. For example, if they standardize $\hat{H}_T(T)$ on T (the focal domain), it is a distance-preserving condition; if they standardize it on Γ (full context space), it is equivalent to applying a consistent $\hat{H}_T(\cdot)$ followed by standardization.

A final concern is that, in many cases, the encoded information from $\hat{H}_T$ may still be a high-dimensional (but structured) object, particularly relative to the effective sample size of the problem at hand. For scalability or predictive performance (such that unimportant, but noisy dimensions do not negatively impact prediction), researchers may further apply some additional dimension reduction technique to the candidate embedding space. For example, in the context of image encoding, Dew (2023) employs Principal Components Analysis (PCA) after VGG-19 image encoding.[13] Ideally, this practice should not lead to violation of the distance-order-preserving assumption.

For our application, we look to embedding models as a means to capture the context of a given unstructured treatment and project it into a latent space.[14] Embedding models are a Generative-AI construct that outputs a dense, numerical vector that represents a text object. The embeddings sit on a latent space comprising a manifold in which similar objects are positioned closer to one another, while objects that differ are positioned further away. Thus, the embedding of a particular text should reside closer to embeddings of other texts with similar content, context, and meaning. A key benefit of this structure is that embeddings reside on a single, common manifold that can be further used for a variety of tasks. Another benefit is that embedding models are generally trained on a collection of tasks that are unrelated to the causal prediction problem at hand, thereby satisfying the requirement of sample-splitting by design. For a more detailed introduction to embeddings, we refer readers to Chernozhukov et al. (2024).

More recently, large language models (LLMs) (a cutting-edge version of embedding models) employ powerful "transformer" architectures that are able to not only capture the context around each word, but also encode the contextual meaning implied by relevant parts of the text that are far apart in a sentence or document (via attention mechanisms). Furthermore, these models are generally trained with massive input from the Internet and are likely to

[13] We also apply PCA in our empirical application.

[14] Please see our discussion in the Web Appendix on other encoding methods and their inability to successfully meet the above criteria.

only improve with time. Therefore, we believe that the embeddings used by LLMs, such as ChatGPT, can better model H_T compared with human codification and other algorithms.

2.3.2 Estimation of G_T

When predicting outcomes or treatment effects for an out-of-sample treatment, researchers need to impose some prior on the mapping (G_T) from the contextual space to treatment effects. This is what makes extrapolation feasible. For example, in the context of combinations of images and text, Bajari et al. (2023) impose a linear functional form on G_T, while in the context of images alone, Dew (2023) assumes a Gaussian process with a squared exponential kernel. Here we make a weaker, more general assumption. In particular, we assume only that G_T is Lipschitz continuous in its arguments, satisfying the following condition:

$$|G_T(H_T(j), w) - G_T(H_T(k), w')| < L \cdot ||(H_T(j), w) - (H_T(k), w')||$$

for any $j, k \in T$ and any $w, w' \in \mathit{supp}(W)$. Here, $L \geq 0$ is the Lipschitz constant, $||\cdot||$ is the Euclidean norm, and $\mathit{supp}(W)$ is the support of W. The above Lipschitz condition restricts how fast the response surface, $G_T(\cdot, \cdot)$, can change with respect to its arguments. This assumption allows us to predict outcomes or treatment effects with flexible partition-based ML algorithms, such as XGBoost (boosted trees) or random forests, which have been shown to have good performance across a wide variety of predictive tasks. To see the role of this assumption, let us consider a partition-based ML algorithm that constructs a partition whose maximum diagram is δ (meaning for any $l \in T$ and any $w \in \mathit{supp}(W)$, we can find a neighbor $j \in T$ and $w' \in \mathit{supp}(W)$ such that $||(H_T(l), w) - (H_T(j), w')|| \leq \delta$), then the Lipschitz continuity assumption implies that the |bias| is bounded by $L \cdot \delta$. By making the partition fine enough ($\delta \to 0$) which, for example, is achieved by adding more trees for boosted trees and random forests, the bias vanishes to zero (a necessary condition for

good prediction performance)[15]. Note that the linear model assumption made by Bajari et al. (2023) which assumes a global functional form of G_T and the Gaussian process (with the squared exponential kernel) assumption made by Dew (2023) which restricts the global variation of G_T are stronger assumptions — it is not hard to verify that both assumptions satisfy the Lipschitz continuity condition.

While the exact approach for constructing prediction intervals depends on the predictive methods researchers employ, such as the standard textbook version of predictive intervals for linear models as in Bajari et al. (2023) and the posterior predictive intervals for Gaussian processes as in Dew (2023), there are two key points to consider. First, inference should take into account the uncertainty arising from employing estimated doubly robust scores but not the uncertainty arising from estimating the nuisance functions (estimated outcome model $\hat{\mu}$ and estimated propensity model $\hat{e}$) used to construct the scores. We need not worry about the estimated nuisance functions because the doubly robust scores satisfy the Neyman orthogonality condition; they "block" the uncertainty induced by the upstream estimations from propagating to the downstream tasks (Newey and McFadden, 1994). Therefore, we can ignore the uncertainty of the estimations before doubly robust scores are constructed and need only consider the uncertainty in the estimated scores and downstream estimates.

However, because these estimated scores are the dependent variable in the predictive stage, robust standard errors (e.g., Eicker–Huber–White standard errors) should be employed to account for the estimation errors in the estimated scores. Second, if $\hat{H}_T$ is estimated on the same dataset for the causal or predictive inference task, then the uncertainty arising from this pre-processing step should also be accounted for. The second point highlights another benefit of using pre-trained models, such as Ada-embeddings from ChatGPT. For a particular dataset, the output of a pre-trained model is deterministic (there is no randomness arising from it). Therefore, we do not need to quantify the uncertainty of $\hat{H}_T$. If one were to instead train $\hat{H}_T$ for the particular task at hand using some candidate ML algorithm, it

[15]Of course this only holds in the asymptotic limit. In practice, there always exists a finite sample bias for every estimator.

is generally computationally costly to quantify the uncertainty arising from this part of the process (e.g., see Bajari et al. (2023)).

Should researchers choose to use flexible nonparametric ML algorithms to model G_T and pre-trained models for H_T, we suggest conducting bootstrapping for constructing predictive distributions/intervals (Breiman, 1996; Fushiki et al., 2005). Bootstrap aggregation essentially works as an ensemble learning technique, which enjoys good predictive performance by strengthening weaker predictors to a strong predictor. More recently, Andrews and Mikusheva (2022) find the bagged prediction smooths discontinuities (if any) in the pre-bagged predictive models. Consequently, the final bagged predictive model often satisfies the Lipschitz continuity assumption more closely compared to the pre-bagged ones.

Finally, one last point to consider is trimming doubly robust scores since the target predicted object is (C)ATE, which is sensitive to outliers. Similar to trimming propensity scores, trimming doubly robust scores can alleviate the impact of outliers, thus improves finite sample predictive performance.

3 Empirical Application - Setting and Data Construction

We illustrate the capabilities and performance, as well as the limitations, of our framework by revisiting and extending the multiple, heterogeneous email treatment application of Ellickson et al. (2023). In that study, the authors employ a doubly robust machine learning estimator to obtain causal effects of a set of email promotions that were sent to 1.3 million customers by an online apparel retailer over 45 days in the spring of 2015. The focal outcomes are customers' engagement and purchase decisions (email opening rates and spending amounts). The heterogeneous treatments considered are 34 unique email subject lines, which the authors hand-encode as multivalued treatment components (categorical encoding). The goal there was to assess the performance of the different subject lines observed in the sample, and

to connect the differential in performance to the multivalued treatment components, i.e., specific keywords and promotion features (e.g., discounts). However, Ellickson et al. (2023) are unable to suggest new creatives (i.e., subject lines of new email promotions) and predict their performance. Therefore, the goal here is predicting the performance of new creatives. In our empirical application, we measure performance based on the purchase amount of an email promotion.

For the purposes of this study, we highlight the variation in outcomes (i.e., purchase amounts) for the heterogeneous treatment effects and the heterogeneity in response to each treatment. In Table 1, we show the various components present in each email subject line and estimated ATE of the purchase amount associated with each email, relative to a common baseline promotion (email 34).[16] As shown, there is considerable variation in both the treatment components of each email, and the performances across the emails. In Figure 2, we showcase the sorted group average treatment effects (GATES) of six emails within our dataset, along with error bars (that are so tight they do not stand out clearly in the graphs). By examining the progression of estimated pair-wise performance of each email with respect to Email 34 (our baseline email), we see that there is considerable heterogeneity across customers in the response to each of these emails. Our extrapolation approach exploits the ability to capture the underlying heterogeneity in treatment mapped from the latent contextual space, which is further moderated by observable customer demographics.

***** Insert Table 1 and Figure 2 about here *****

Ellickson et al. (2023) provide a variety of tests and evidence indicating that the firm engaged in a sufficiently randomized targeting of their emails to ensure proper estimation of effects that verify the identification requirements of Assumptions 3 and 4. In particular, they establish that the variables used for targeting are fully observed and provide strong evidence of residual randomization (overlap). The pre-treatment covariates used for targeting fall into

[16]Please see Section 4.1 for a discussion on how we estimate the doubly robust scores that are then aggregated into the ATEs reported in Table 1 and GATEs shown in Figure 2.

one of three categories: demographics, RFM variables, and engagement variables. In what follows, we use these variables to both generate the response surface, and for developing policy suggestions for our latent treatments. Table 2 provides a list of these customer characteristics and their summary statistics.

***** Insert Table 2 about here *****

4 Extrapolation Method Evaluation

To assess the ability of our framework to properly quantify the effects of novel heterogeneous treatments, in terms of both aggregate and heterogeneous effects, we proceed in the following stages. We begin by following the process described in section 2.2 to obtain estimates of the doubly robust scores. Specifically, we perform a series of leave-one-out (LOO) exercises to compare the predictions of our estimation framework against a strong proxy for the ground truth, the doubly robust scores of each email promotion. Our first validation exercise is to compare the extrapolated average treatment effects (ATE_{Exp}) (predicted average performance) to the estimated average treatment effects computed from the doubly robust scores (ATE_D), which captures the actual performance. Our second validation exercise takes a more granular approach by comparing the sorted average treatment effects (GATES) of our prediction scores to the GATES-analogs of the doubly robust scores (conditional averages). Our third and final validation exercise is a comparison of policy performance between using the extrapolation scores (ITE_{Exp}) and the true doubly robust scores (ITE_D) via a data-adaptive policy algorithm (policy tree).

4.1 Generating Doubly Robust Scores

The procedure to generate the scores that represent the ground truth closely follows the approach used in Ellickson et al. (2023), of which the key steps are discussed in section 2.2. To obtain these pairwise doubly robust scores, we must first estimate the nuisance

functions $\mu(\cdot)$ (outcome model) and $e(\cdot)$ (propensity score model). Both tasks are performed using random forests. This method has been shown to perform well in a variety of causal domains in which interactions and feature selection are needed (Chernozhukov et al., 2024). Following standard practice for approaches that involve propensity weighting, we further trim the estimated propensity scores, $\hat{e}$, such that they lie in between 0.01 and 0.99 to avoid bias caused by outliers (Crump et al., 2009), then we construct the estimated pairwise doubly robust scores, $\hat{\tau}_{jk,dr}$, following equation 1. For the purpose of producing good finite-sample predictive performance downstream, we also trim the obtained doubly robust scores by using the thresholds of 1% and 99%.

4.2 Generating $\hat{H}_T$

To estimate H_T we use the Ada-0002 engine that powers ChatGPT as our contextual embedding model. The Ada engine translates text into a dense numeric vector of 1,536 features.[17] As noted earlier, given sample constraints, we require an additional dimension reduction beyond the full embeddings. In particular, we follow Dew (2023) in using PCA reduction. We, ultimately, select 28 dimensions, which capture over 99% of the variance in the embeddings.[18] Figure 3 shows the embedding component plots for our six example emails, and Figure 4 shows their PCA-reduced counterparts. In both figures (particularly in Figure 4), there are clear commonalities between emails that share smaller distances. We then test if the distance order between treatments is maintained between the raw embeddings and the PCA-reduced embeddings, relying on the calculation of their Euclidean distance. In the manifold, Euclidean distance can be used by the researcher to view the similarity of their semantic meaning in the embedding space.[19] Figure 5 is a heat map of the Euclidean

[17]The API of Ada-0002 engine offers an option to choose how many features to output. We recommend to output all features. If researchers want to further reduce the dimension, they can do so themselves downstream. In doing so, they can then check whether the dimension reduction violates the distance-order-preserving assumption.

[18]We illustrate the performance of each level of dimension reduction through PCA in Figure A.6.

[19]A common practice to measure distance in the embedding space in the ML literature is to use cosine similarity. However, cosine similarity has been recently challenged as a valid metric to compare similar

distance between the embeddings of our 34 email subject lines, for both the raw contextual embeddings and the PCA reduced embeddings. The lighter regions show greater distance between email subject lines while the darker regions show email subject lines that are closer in the contextual manifold. Comparing the heat maps between the raw embeddings (left) and PCA-reduced embeddings (right), we see only minor differences in the values. More importantly, the ranking of distances between emails is maintained, which is a requirement for extrapolation from the context space into novel heterogeneous treatments — the PCA estimates serve as an adequate proxy for the Ada embeddings in this exercise. Thus, in what follows, we will work only with PCA estimates, while continuing to refer to them as Ada embeddings for continuity.

***** Insert Figures 3, 4, 5 about here *****

4.3 Generating $\hat{G}_T$ and Predicting Effects of New Treatments

Having obtained $\hat{H}_T$ for the different encoders, we proceed to estimate $G_T(\cdot)$ for each encoder and predict the ATEs of new treatments (ATE_{Exp}). To simulate the scenario of a new treatment and evaluate the predictive performance, we conduct a series of LOO prediction exercises across our email treatments. Our goal is to predict the performance (treatment effect) of a given focal email (k_{focal}), so we generate $\hat{G}_T(\cdot)$ on the other 32 emails, leaving out the focal email, and then use $\hat{G}_T(\hat{H}_T(k_{\text{focal}}))$ to predict the relative performance of the focal email. [20] We treat the estimated ATE of the focal email as the ground truth, compare how close the ground truth (ATE_D) is to the predicted effect (ATE_{EXP}). Specifically, the "ground truth" is the average of the doubly robust scores in the held out data of the focal email (i.e., the actual performance).

pieces of text in the embedding space Steck et al. (2024). We suggest using Euclidean distance instead. For completeness, we provide cosine similarity plots in Section A.2 of the Web Appendix.

[20]Note, although we have 34 emails in the dataset, Email 34 is considered the baseline email. Therefore, ATEs of 33 emails are calculated with respect to Email 34. In the LOO exercise, each time we leave out an email for evaluating predictive performance, we are training on the ATEs of the remaining 32 emails.

When implementing the predictive exercises, we employ an optimally tuned XGBoost algorithm (please refer to Section A.3 for more details). We chose XGBoost as its performance in prediction tasks is well known, and the boosted tree structure allows for "automatic" interactions between features through recursive splitting. Further, XGBoost also provides automatic feature selection, which is useful in our application given the high-dimensional nature of the PCA-reduced, contextual embeddings. This adaptive feature enhances the successful predicting the outcomes of new emails, which is especially useful since we only have 32 email contrasts for estimating $G_T(\cdot)$.[21]

4.4 Aggregate-Level Comparison

To assess the performance of our framework's ability to predict the outcomes of novel heterogeneous treatments (new emails), we compare the 95% confidence interval of the predicted outcome of the focal email (extrapolated performance) versus the confidence interval of the doubly robust scores (ground truth). Figure 6 presents the performance of our method in generating predictions for this exercise. Only two of the emails fail to have overlapping confidence intervals, though one email has a large confidence interval and noticeably different outcome, Email 5. To further check the predictive power of our model, we regress the ATE_D of each email on the corresponding ATE_{Exp}. The goal is to quantify how much variation our extrapolated scores capture from the actual, observed scores. Using this approach, we find the R^2 to be 78.6%, providing further evidence that our method yields meaningful estimates of the outcomes from novel heterogeneous treatments.

***** Insert Figure 6 about here *****

We note there are a few clear and revealing cases in which we fail to accurately capture the ground-truth. In what follows, we use Euclidean distance to illustrate the underlying

[21] In general, researchers should select a predictive model that is suitable for their context, for example, Bajari et al. (2023) select a linear model in their embeddings, which is motivated by hedonic theory from economics.

issues. Recall that Figure 5 provides a heatmap of the Euclidean distance between each focal email and the other 32 pair-wise contrasts used to build the model. Examining the Euclidean distance measures in the heatmap, we find a consistent pattern based upon the treatments that show confidence intervals closer to the doubly robust scores (ground truth) versus those that do not. Emails that share a lower Euclidean distance (more similar) with other emails yield better performance. Comparing the Euclidean distance measures for Email 5 (a simple Happy Birthday email) with the other 32 emails in our dataset, we note the following patterns. First, the lowest Euclidean distance between Email 5 and those used to build the predictive model for Email 5 is .67 — this lack of similarity is the primary contributing factor to the poor performance. This should not be surprising because XGBoost is a partition-based estimation technique that heavily relies on local information. Second, the emails that are shown as similar do not share the same context with the focal email. For example, one of the similar emails wishes the receipt a Happy Memorial Day. While the positive sentiment is shown in both emails, the contexts are clearly divergent. Since a researcher cannot directly interpret the embedding structure, this exercise is important to describe both the features and context of the treatment object.

In contrast to the Happy Birthday email, one email that shows clear congruence between the ATE_{Exp} and ATE_D is Email 13, which offers 30% off Clearance Items. The Euclidean distances of the top four similar emails with Email 13 are all below .36. Exploring the contextual nature of these emails, we find great similarity. Specifically, the emails within the top four are clearance emails with differing percentage off discounts. The closest email to Email 13 is a 20% off clearance items email, with another 40% off clearance email (but with a slightly different subject line) as the next highest. The closeness in terms of both the context (clearance items) and ordering of discounts shows that Ada embeddings can be used to describe relevant email characteristics. Overall, to predict the effectiveness of novel heterogeneous treatments, there must be support on the observed set of contextual features (experimental design points) around the treatments for prediction. Researchers and

managers must consider whether the feature space shows similar richness before they perform predictive tasks. [22]

4.5 Heterogeneity in Response Comparison

Our prior analysis revealed that our proposed framework yields accurate estimation of ATEs, which suggests that the contextual embeddings are able to capture the signal provided by the latent contextual structure of the promotions. However, average performance does not reveal if our method is able to account for the heterogeneous reactions to novel treatments. For use in targeting, we must account for the moderating effect of customer characteristics on the treatment outcomes. To explore the predictive power of our method at this more granular level, we use sorted average treatment effects (GATES). Proposed in Chernozhukov et al. (2017, 2018), GATES orders the scores in ascending order and then partitions them into subgroups. GATES analysis both eliminates some idiosyncratic noise from the doubly robust scores and provides a distribution of effects for later analysis. For example, in Chernozhukov et al. (2018), after examining the GATES, they then look to identify key demographic characteristics that are associated with the positive versus negative effects via classification analysis (CLAN). For our purposes, in assessing out-of-sample predictive performance, we use the GATES to separately bin ITE_{Exp} and ITE_D to see if our method effectively recovers heterogeneous effects at a more granular level.[23]

For each email in our LOO exercise, we create 1000 bins (representative of .1% of the distribution) based on the sorted ITE_D scores. Within these bins, we compute the associated conditional average treatment effect from both the doubly robust scores ($GATES_D$) and the extrapolated scores ($GATES_{Exp}$). Next, for each email, we regress the $GATES_D$ on the

[22] We also considered a number of other encoders for the task of computing the ATEs in our application. We have included them and a discussion of their performance in the Web Appendix. In brief, the use of PCA-reduced, contextual embeddings performs better than other candidate methods.

[23] For completeness, Figure A.7 presents the ATE_D versus the extrapolated ATE_{Exp} when we include customer demographics. Noticeably, the error bars for the prediction are larger when customer demographics are considered, which is a direct cause of lack of support in some regions of customer characteristics in the dataset. While this may be a cause for concern, our analysis of the GATES provide evidence that we are able to effectively recover the trajectory of heterogeneous scores of novel heterogeneous treatment.

$GATES_{Exp}$ and compute the R-squared to quantify the amount of actual variation in the scores captured by our projection. Across the 33 email experiment, our average R-squared is 65.36% (Median R^2: 66.43%), with a $10\% - 90\%$ interpercentile range of 44.86% to 82.63%. Given that heterogeneous effects are constructs that are much harder to recover, compared to ATEs, the predictive power shown in this exercise supports the success of our approach with only 33 treatments in our dataset.

To demonstrate how the GATES can be directly tied to usable, customer-level features, we perform the following classification analysis on one email. For Email 13, which we have highlighted in prior sections, we compute a set of RFM variables based upon the median split of the GATES. Recall that GATES are constructed based upon the ascending order of the doubly robust scores, thus we are interested in seeing if there is systematic difference in customer-level characteristics between those with low scores versus high scores. Table 3 shows the average profile of customers based upon a median split of the GATES. Those customers associated with a higher GATES (above the median) show higher values in recency-frequency-moneytary (RFM) variables than their counterparts with lower GATES (below the median). Specifically, the customers with higher ITE_D open emails more frequently, they have purchased more recently and frequently, and have higher purchase amounts. Thus, we argue, the use of the GATES directly links to known behavior of customer characteristics through RFM variables.

***** Insert Table 3 about here *****

4.6 Policy Recommendation Comparison

Our prior exercises examine if our method can accurately predict both the average effect of new treatments, and their heterogeneous effects. Given that practitioners choosing to deploy treatments seek to do so in the most efficient manner, we now explore the use of a data-driven targeting policy based upon the predictions of our model. Specifically, we compare the relative accuracy of assigning treatment using ITE_{Exp} for a novel treatment against the

corresponding treatment assignments using ITE_D. In other words, how does the predictive approach compare to knowing the actual performance in the field (i.e., ex ante rather than ex post).

For our data-driven targeting policy, we employ the recursive partitioning method proposed by Athey and Wager (2021), also known as policy tree. The objective of this algorithm is to find the best set of heuristics for allocating the treatments under consideration. We implement this assignment using the doubly robust scores. Treating these scores as outcome signals, the algorithm explores each variable in sequence to balance the gains of allocating a given treatment to every observation in that partition against the regret of misallocation. The result is a decision tree that provides the heuristics associated with treatment assignment. Thus, for any set of characteristics, the user of policy tree can obtain a prediction of the treatment assignment.

For our exercise, we contrast the assignments proposed by a policy tree of depth 2 against those of the Oracle policy assignment. Oracle policy assignment assumes that, for each observation, a given heuristic rule is applied. In our context, this heuristic dictates that if a customer's doubly robust score for an email, relative to our baseline Email 34, exceeds 0, the customer receives the focal email; otherwise, the customer is assigned the baseline email. Since our doubly robust scores are pairwise contrasts between a focal email and email 34, this exercise approximates the targeting of a chosen email against email 34 (the baseline email). Thus, we calculate the number of correct predictions between these two emails using policy tree versus the Oracle prediction to measure the percentage of correct predictions. We then compute the ratio of correct predictions between the policy tree created with extrapolated scores PT_{Exp} against the policy tree created with doubly robust scores PT_D.

For illustrative purposes, we showcase the six emails used in our motivating example to highlight the performance of our method against the actual doubly robust scores in driving the correct policy suggestions. The ratio of correct predictions between PT_{Exp} and PT_D for our six example emails, on average, is 0.82. Of these six emails, the lowest value of the ratio

of correct prediction is 0.68 and the highest is 0.96. For these six emails, we compute the ratio of revenue from targeting based on PT_{Exp} versus PT_D. To compute revenue, we sum ITE_D based on the policy selection from PT_{Exp} and PT_D, respectively. We find that the policy associated with scores extrapolated from our method recovers 71.72% of the revenue obtained when basing the assignments on actual performance.

Based upon these results, we conclude that our predictive method is able to adequately capture the performance of novel heterogeneous treatments. We turn next to an empirical exercise in which Generative AI is used to both develop, and assess the performance of, new email creatives.

5 Novel Treatment Creation and Evaluation using ChatGPT

Our proposed method relies on Ada embeddings to accurately represent the latent context. They are then used to map the response surface from treatment to outcome. These embeddings are also the foundation of ChatGPT's generative model. To conclude the paper, we examine one further use case: evaluating ChatGPT's ability to generate new emails and accurately predict their performance. To assess ChatGPT's ability to predict the success of suggested interventions, we use our method to calculate the extrapolated scores for a random subset of individuals. With these scores, we can then compare the ranking provided by ChatGPT against the predicted performance provided by our model.

We conduct our exercise as follows. First, we provide ChatGPT with a set of features and contextual language it can use to generate new emails. Second, we further define that the generated content must be an email subject line and also constrain the number of characters that ChatGPT can use to create the subject line. This step is to ensure the generated content is comparable to content (subject line of an email) observed in the data.[24] Third,

[24]Please see our discussion in the web appendix for the actual prompting used in ChatGPT.

we ask ChatGPT to create 5 new email promotions and then rank them from best to worst in its assessment of generating incremental revenue. Fourth, we assess the performance of the generated content by utilizing our complete set of average treatment effects of 33 emails relative to our baseline email 34 to construct an XGBoost model, as outlined in our previous sections. Table 4 shows the generated emails, the ranking provided by ChatGPT, and the extrapolated ATEs of these emails using our method.

***** Insert Table 4 about here *****

Two findings immediately stand out: (1) ChatGPT is indeed able to suggest coherent email subject lines, some of which are predicted to perform very well; (2) ChatGPT is unable to determine the relative success of different emails.[25] Based on our findings, we conclude that ChatGPT and other large generative models, though trained on an extensive text corpus (so they can estimate $H_T(\cdot)$ well), lack the specialized domain knowledge to predict performance. In our case, the specialized knowledge is the estimated response surface, $G_T(\cdot)$ function. By combining the latent contextual information ($\hat{H}_T(\cdot)$) with observable outcome data to estimate the $G_T(\cdot)$ function, we are able to combine domain specific knowledge in the creation of an effective tool to predict the performance of novel heterogeneous treatments.

5.1 Targeting of Novel Treatments

In our prior analysis, we showed that the use of ITE_{Exp} to generate a policy tree can have similar performance as a policy tree grown with ITE_D. Based upon this knowledge, we further posit that a manager using ChatGPT to create new content can use our approach to not only see how well each suggestion does on average, but also prescribe personalized targeting decisions. We use our model with all 33 pair-wise email ATEs to predict the ITE

[25]For completeness, we also ask ChatGPT to rank order the actual emails we observe in our empirical setting. We simulate the responses 30 times and calculate the average ranking of them based upon ChatGPT and present the differential between actual performance versus ChatGPT's suggested performance in the Web Appendix in FigureA.8. This exercise show lack of meaningful assessment by ChatGPT to our example with generated content.

for 10,000 simulated individuals for each of two generated emails. For this exercise, we use two GPT generated emails with relatively similar predicted average performance: Limited Time Offer: 50% Off Shipping on All Products! (Opt 1) and Discover New Products: Free Returns on Every Order! (Opt 2). One email's focus is on providing an immediate shipping discount, while the other offers free returns. The difference in promotional type should provide an interesting contrast for potential targeting implications. For our subset of simulated individuals, the ATE of Opt 1 is $-.068$, while the ATE of Opt 2 is $-.071$.[26] Our goal here is to see if there is a possibility of generating a higher lift through careful, data-driven targeting.

To see if there are gains from targeting, we estimate a shallow, two-level tree using the policy tree algorithm outlined in Section 4.6. Figure 7 presents the decision tree of this targeting exercise. As shown, the three most critical variables in targeting the emails are the total number of orders placed by the individuals in the past, their web spending, and the recency of their purchases. Individuals with low amounts of orders (less than 10) and low web spending (average web spending less than \$47) should get Opt 1, while greater web spending results in Opt 2 being assigned. If the individual has a greater order frequency (more than 10 orders), then the context of targeting turns upon how recently has the individual purchased. Someone who has purchased very recently should get Opt 2, while all others in the high order frequency category should get Opt 1. Overall, the recommendations of the policy tree are intuitive and reflect the patterns revealed in the earlier CLAN exercise. Those individuals who have engaged with the firm less (in both frequency and purchase amount) likely need an upfront discount on shipping to be willing to purchase. While those that purchase less often, but in higher amounts, would benefit from free returns. On the other hand, frequent purchasers, who engage with the company more often, would get greater benefit from the free returns. The sum total of this targeting policy results in an overall lift of -.04, which is an improvement of approximately 40%.

[26]The negative numbers of the ATE are not indicative of losing money per se. Instead, they show that these emails perform worse in revenue generation than our baseline email.

***** Insert Figure 7 about here *****

6 Conclusion

In this paper, we propose a framework that allows for a researcher to extrapolate the performance of a novel treatment using contextual embeddings and machine learning nested in the potential outcomes framework. We show that contextual embeddings summarize the latent context of unstructured data seen in digital marketing, a common marketing application. Further, since contextual embeddings reside on a common linguistic manifold, we show that our method is able to predict the outcomes of novel heterogeneous treatments. By linking the latent context with customer characteristics and doubly robust scores, we generate a response surface that measures the effects of novel treatments. We validate our method by exploring the performance at the ATE, GATES, and policy recommendation levels. Working with only 32 ATEs in total, we show consistent recovery of heterogeneous effects and optimal policy choices for treatments not included in the trained model.

Beyond our proposed framework, our study highlights the usefulness of using contextual embeddings directly in economic applications. The ability to quickly and inexpensively obtain these embeddings highlights their value as high-dimensional treatment objects within a latent class estimator. Though they have been used as confounders in other studies (i.e., Bajari et al. (2023)), operationalizing them directly is an important new step in the empirical literature. Future research could explore their direct inclusion and improve interpretability to refine estimates of causal objects of interest. Additionally, given that other media types such as video and images can also be operationalized as embeddings, our latent class causal framework can be readily extended to various marketing vehicles. Moreover, we emphasize that while AI offers valuable insights, it cannot replace human judgment. We provide a way for a marketing manager or researcher to judge the choices made by ChatGPT by providing a framework that allows for off-policy evaluation of heterogeneous treatments. Last, our

study demonstrates the viability of ChatGPT's embeddings in uncovering language structures previously inaccessible to academic research, opening up new avenues of investigation in marketing as marketers increasingly engage with customers through text-based initiatives.

References

Andrews, I. and A. Mikusheva (2022). Gmm is inadmissible under weak identification. *arXiv preprint arXiv:2204.12462*.

Angelopoulos, P., K. Lee, and S. Misra (2024). Value aligned large language models. *SSRN Electronic Journal*.

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research 55*(1), 80–98.

Athey, S. and S. Wager (2021). Policy learning with observational data. *Econometrica 89*(1), 133–161.

Bajari, P., Z. Cen, V. Chernozhukov, M. Manukonda, S. Vijaykumar, J. Wang, R. Huerta, J. Li, L. Leng, G. Monokroussos, et al. (2023). Hedonic prices and quality adjusted price indices powered by ai. *arXiv preprint arXiv:2305.00044*.

Balakrishnan, R. and R. Parekh (2014). Learning to predict subject-line opens for large-scale email marketing. In *2014 IEEE International Conference on Big Data (Big Data)*, pp. 579–584. IEEE.

Breiman, L. (1996). Bagging predictors. *Machine learning 24*, 123–140.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernandez-Val (2017). Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research.

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

Chernozhukov, V., C. Hansen, N. Kallus, M. Spindler, and V. Syrgkanis (2024). Applied causal inference powered by ml and ai. *https://causalml-book.org*.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*(1), 187–199.

da Costa, L. S., I. L. Oliveira, and R. Fileto (2023). Text classification using embeddings: a survey. *Knowledge and Information Systems 65*(7), 2761–2803.

Dew, R. (2023). Adaptive preference measurement with unstructured data. *Available at SSRN 4641773*.

Dubé, J.-P. and S. Misra (2023). Personalized pricing and consumer welfare. *Journal of Political Economy 131*(1), 131–189.

Egami, N., C. J. Fong, J. Grimmer, M. E. Roberts, and B. M. Stewart (2022). How to make causal inferences using texts. *Science Advances 8*(42), eabg2652.

Ellickson, P. B., W. Kar, and J. C. Reeder (2023). Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science 42*(4), 704–728.

Fong, C. and J. Grimmer (2023). Causal inference with latent treatments. *American Journal of Political Science 67*(2), 374–389.

Fushiki, T., F. Komaki, and K. Aihara (2005). Nonparametric bootstrap prediction. *Bernoulli 11*(2), 293–307.

Gekhman, Z., G. Yona, R. Aharoni, M. Eyal, A. Feder, R. Reichart, and J. Herzig (2024). Does fine-tuning llms on new knowledge encourage hallucinations?

Girotra, K., L. Meincke, C. Terwiesch, and K. T. Ulrich (2023). Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available at SSRN 4526071*.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, Part I: Causal models, structural models and econometric policy evaluation. *Handbook of Econometrics 6*, 4779–4874.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jaidka, K., T. Goyal, and N. Chhaya (2018). Predicting email and article clickthroughs with domain-adaptive language models. In *Proceedings of the 10th ACM Conference on web science*, pp. 177–184.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung (2023, mar). Survey of hallucination in natural language generation. *ACM Comput. Surv. 55*(12).

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

Nguyen, N., J. Johnson, and M. Tsiros (0). Unlimited testing: Let's test your emails with ai. *Marketing Science 0*(0), null.

Reisenbichler, M., T. Reutterer, D. A. Schweidel, and D. Dan (2022). Frontiers: Supporting content marketing with natural language generation. *Marketing Science 41*(3), 441–452.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*(429), 122–129.

Steck, H., C. Ekanadham, and N. Kallus (2024). Is cosine-similarity of embeddings really about similarity? *ACM Web Conference 2024 (WWW 2024 Companion)*.

Yoganarasimhan, H., E. Barzegary, and A. Pani (2020). Design and evaluation of personalized free trials. Technical report, University of Washington.

Tables and Figures

		Category		Promotion Characteristics						Semantic Characteristics								
Email	Characters	Product	Clearance	Discount	Cash	Free Gift	Free Shipping	50% Off Shipping	Free Returns	Personalized	Miscellaneous	SC 1	SC 2	SC 3	SC 4	SC 5	SC 6	ATE
1	60	0	0	0.25	0	0	0	0	0	0	0	1	0	0	0	0	1	-0.0410
2	32	1	0	0.35	0	0	0	0	0	0	0	1	0	1	0	0	0	-0.0170
3	32	1	0	0.4	0	0	0	0	0	0	0	1	0	1	0	0	0	-0.0097
4	63	0	0	0.3	0	0	1	0	0	0	0	0	0	0	0	0	0	-0.0537
5	14	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0.3689
6	50	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	-0.0332
7	60	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0.0002
8	35	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0.0056
9	58	0	0	0.45	0	0	0	0	0	1	0	1	0	1	1	0	1	0.0131
10	50	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0.0098
11	76	1	0	0.35	0	0	0	0	0	0	0	0	0	1	0	1	1	0.0006
12	59	0	1	0.2	0	0	0	0	0	0	0	1	0	1	0	0	1	-0.0264
13	59	0	1	0.3	0	0	0	0	0	0	0	1	0	1	0	0	1	0.0683
14	65	1	0	0	0	0	1	0	1	0	0	0	0	1	0	1	1	-0.0246
15	65	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	-0.0370
16	65	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	-0.0081
17	69	0	1	0.3	0	0	0	0	0	0	0	1	0	1	0	0	1	0.0992
18	69	0	1	0.4	0	0	0	0	0	0	0	1	0	1	0	0	1	0.0771
19	71	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	1	-0.0308
20	71	1	0	0.35	0	0	1	0	0	0	0	0	0	0	0	0	1	-0.0143
21	70	0	0	0.55	0	0	1	0	0	0	0	0	0	1	0	1	1	0.0017
22	64	1	0	0.5	0	1	1	0	1	0	0	0	0	1	0	0	1	-0.0207
23	56	0	0	0.45	0	0	0	0	0	0	1	0	0	1	0	0	1	-0.0113
24	67	0	0	0.35	0	0	0	0	0	0	0	1	0	0	0	0	1	-0.0164
25	65	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	-0.0365
26	56	1	0	0.5	0	0	0	0	0	0	0	1	0	1	0	0	1	-0.0439
27	55	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0.0009
28	63	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0.0267
29	69	0	0	0.7	0	0	1	0	0	0	0	0	0	1	0	0	1	-0.0226
30	59	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	1	-0.0194
31	59	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	-0.0074
32	72	1	0	0.1	0	0	0	1	0	0	0	0	0	1	0	0	1	0.0750
33	69	1	0	0.8	0	1	0	0	0	0	0	0	0	0	0	0	1	-0.0326
34	66	0	0	0.4	0	0	0	0	1	0	0	1	0	1	0	0	1	NA

Table 1: **Descriptive Representation of the 34 Email Subject Lines in our Dataset.** This table summarizes subjectline characteristics (i.e., human-coded treatment components) for our 34 email promotions across three categories: merchandise, promotion, and semantic characteristics. Binary indicators show if a characteristic is present in the subject line. For confidentiality, some discounts and character counts are altered, and six semantic characteristics are anonymized. The last column shows the ATE of each email compared to the baseline (Email 34). Negative ATE numbers indicate that these emails generate less revenue than the baseline, not that they result in a loss.

Customer Characteristics	Pretreatment Covariates		Summary Statistics	
	Variable	Description	Mean	SD
Recency	Days_Reg	Days since the recipient registered with the firm to receive emails	1595.08	1288.81
	Days_Open	Days since the recipient opened her last email	67.33	126.29
	Days_Click	Days since the recipient clicked on her last email	284.97	558.00
	Days_Pur	Days since the recipient made a purchase from her last email	475.63	700.19
Frequency	Order_Count	Number of purchases made by the recipient in the past two years.	3.36	5.05
	Tot_Dept	Total number of departments that the recipient has shopped from	4.25	4.19
Monetary	Ave_Ret_Spend	Average $ amount spent each time the recipient makes a retail purchase	10.81	30.21
	Ave_Web_Spend	Average $ amount spent each time the recipient places an order online	42.91	45.14
Habitual	Catl_Book	True (1) if the recipient received catalog book	0.11	0.32
	Custom_Choice	True (1) if the recipient customized buying preferences	0.23	0.36
	Pur_Off	True (1) if the recipient makes retail purchase	0.43	0.46
	Pur_On	True (1) if the recipient makes online purchase	0.89	0.78
Demographic	Bday	Birthday Month of the recipient	6.26	3.84
	Age	Age bracket of the recipient	3.57	1.51
	Income	Income bracket of the recipient	4.74	2.17

Table 2: **Customer-Level Heterogeneity Variables.** This table summarizes the 15 variables used by the focal firm to describe and target their customers with email promotions. These variables are grouped into categories based upon RFM variables, the shopping habits, and demographic characteristics. We provide a description of each variable and summary statistics. These variables are used to generate the targeting policy of the newly developed email creatives from ChatGPT. Note that this table replicates Table 3 in Ellickson et al. (2023).

	Customer Characteristics				
Group	AOV Retail ($)	AOV Web ($)	Days Open	Days Purchase	Order Count
Below Median	8.79 (28.89)	16.79 (34.51)	195.69 (284.71)	953.39 (875.78)	0.69 (1.64)
Above Median	9.77 (27.76)	51.48 (46.81)	92.25 (173.82)	380.07 (574.21)	3.85 (4.65)
Overall Average	9.28 (28.33)	34.14 (44.63)	143.97 (241.48)	666.73 (794.06)	2.27 (3.83)

Table 3: **Example of Customer Characteristics Split for Email 13 Based on Sorted Average Treatment Effects.** This table shows the demographic differences based upon a median split of the sorted average treatment effects (GATES) from Email 13's customers. The GATES are ordered from lowest to highest based on the doubly-robust scores calculated by DML, thus GATES below the median are lower than those above the median. We show the average customer characteristics (with standard error in parentheses) based on this median split of the GATES in the table. As shown, those customers associated with a higher GATES are more engaged with the company, they have higher spend on both retail and website. In addition, they are more engaged with opening emails (lower Days Open), purchased more recently (lower Days Purchase), and purchased more often (higher Order Count). Thus, the GATES can be thought of as a proxy for the known effect of RFM variables.

Subject Line	ATE	GPT Ranking
Exclusive Product Sale: Extra 20% Off Today!	0.043	3
Other Kinds Extravaganza: Code for Extra Savings!	-0.010	4
Clearance Event: BOGO Deals + 30% Off Products!	-0.023	1
Discover New Products: Free Returns on Every Order!	-0.028	5
Limited Time Offer: 50% Off Shipping on All Products!	-0.089	2

Table 4: **Extrapolated Performance of GPT Generated Emails with GPT Rankings.** This table shows the predicted average performance (ATE) of each GPT generated email. Emails are ranked based upon the ATE performance, from best to worst. In addition, we include a column that shows the rankings of each email's performance based upon GPT's ranking. ATEs are calculated from an optimally-tuned, XGBoost model with the PCA-reduced, Ada Embeddings as our set of feature variables and the dependent variable is the doubly robust scores from our 33 email pair-wise contrasts. After building the XGBoost model, we use the PCA-reduced, Ada Embeddings from the new email creatives to predict their performance. As shown, GPT is useful for ideation but is unable to properly rank order the emails based upon their performance.

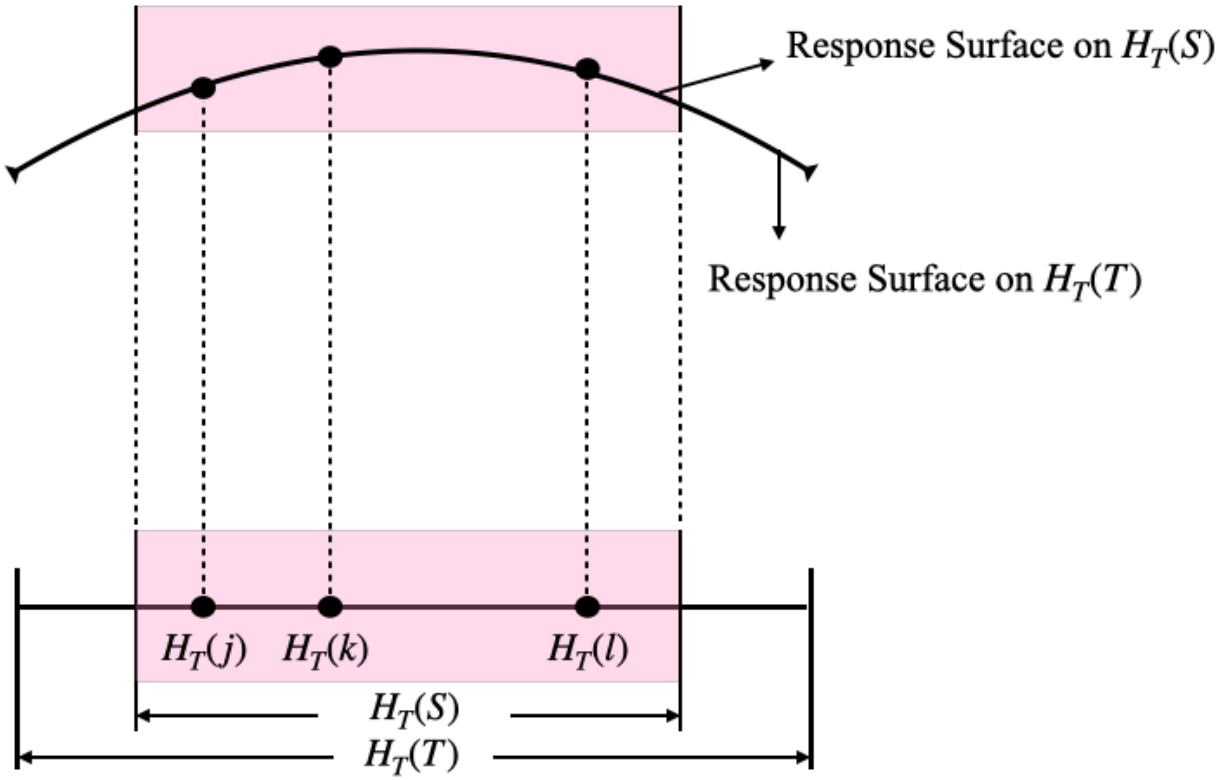
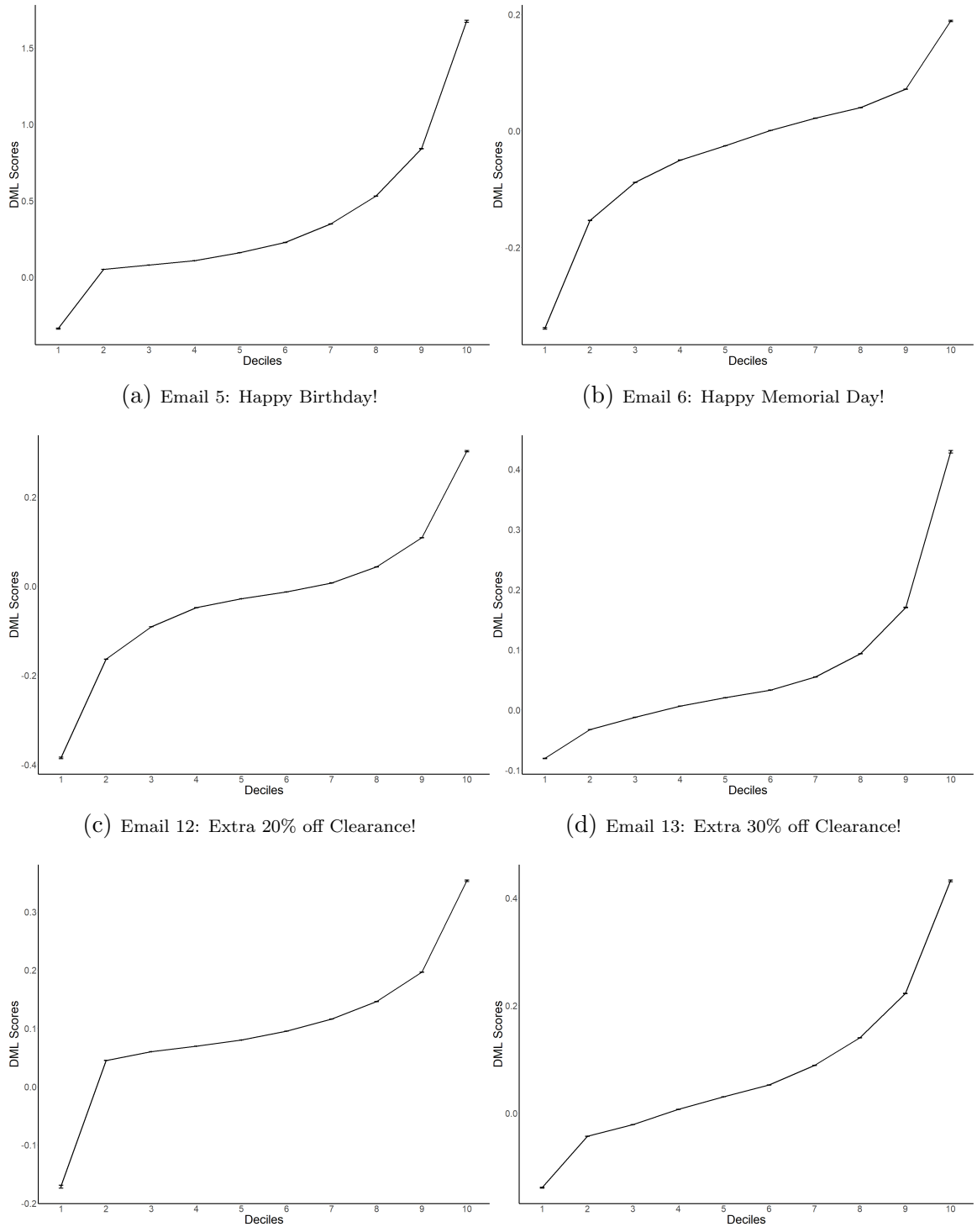


Figure 1: **Counterfactual Prediction.** In this figure, we show the intuition behind our latent treatment framework. The treatment assignment rule, $D(\cdot)$, samples all points with different propensities in the observed treatment set, S, e.g., j, k and l. Based on our assumptions, we can locate the treatment response surface on context set observed in data, $H_T(S)$. After imposing some choice of functional form assumption on the response surface, we can extrapolate the response surface to the context set of interest, $H_T(T)$.



(a) Email 5: Happy Birthday!

(b) Email 6: Happy Memorial Day!

(c) Email 12: Extra 20% off Clearance!

(d) Email 13: Extra 30% off Clearance!

(e) Email 17: Extra 30% off Clearance - Online Only!

(f) Email 18: Extra 40% off Clearance - Online Only!

Figure 2: **GATES of Six Example Emails.** We showcase the sorted average treatment effects (GATES) of six emails within our dataset, along with error bars (that are so tight they do not stand out clearly in the graphs). ITEs of each email are assigned to ten groups in ascending order.

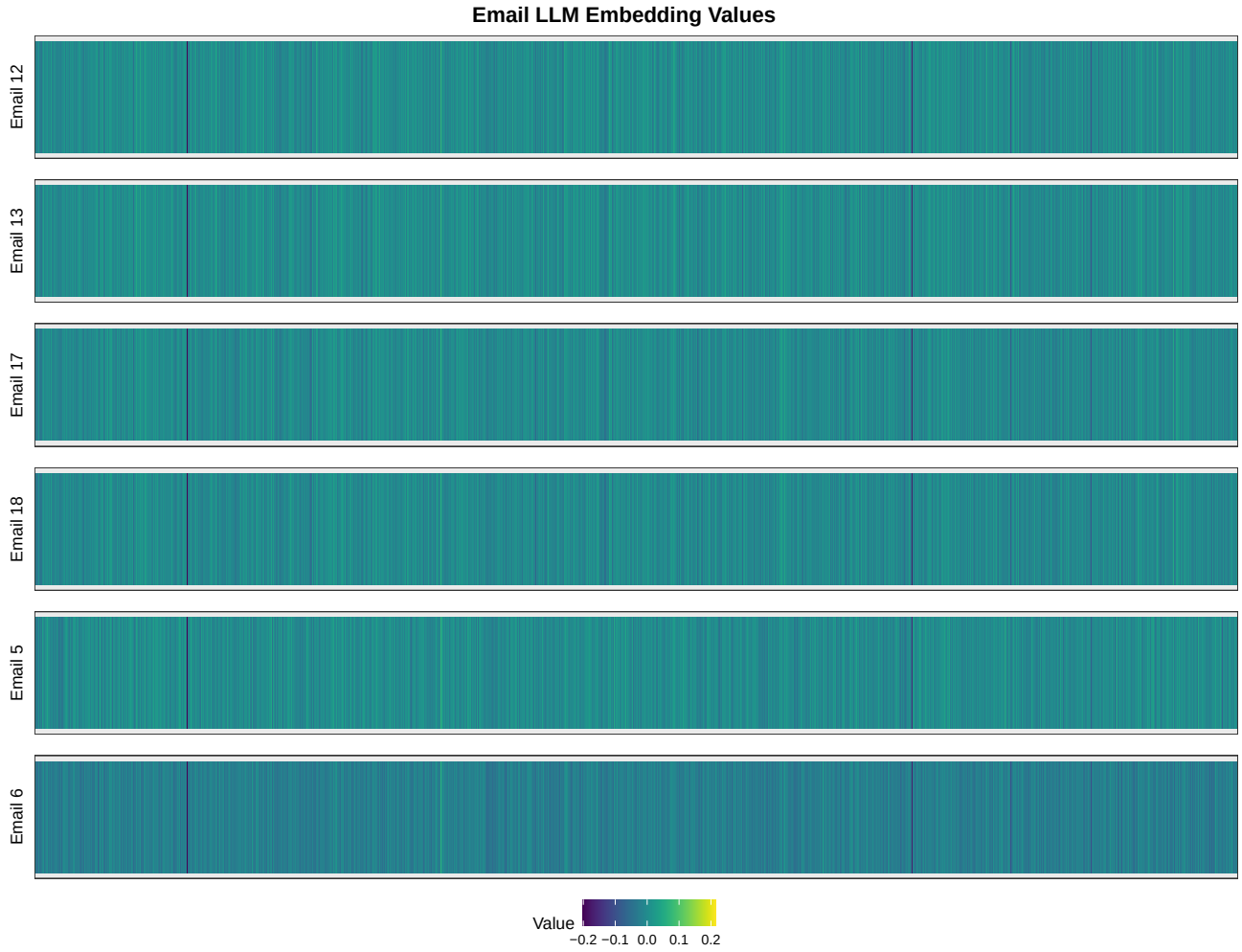


Figure 3: **Ada Embeddings of Six Example Emails.** This figure shows the Ada Embeddings of 6 example emails. In this figure, Email 12 has subject line "Extra 20% off Clearance!" Email 13 has a subject line "Extra 30% off Clearance!" Email 17 is an "Extra 30% off Clearance - Online Only!" Email 18 is an "Extra 40% off Clearance - Online Only!" Email 5 is a "Happy Birthday!" email. Email 6 is a "Happy Memorial Day!" email. As shown, there is commonality between the emails that show the lowest Euclidean distance, such as Emails 12 and 13.

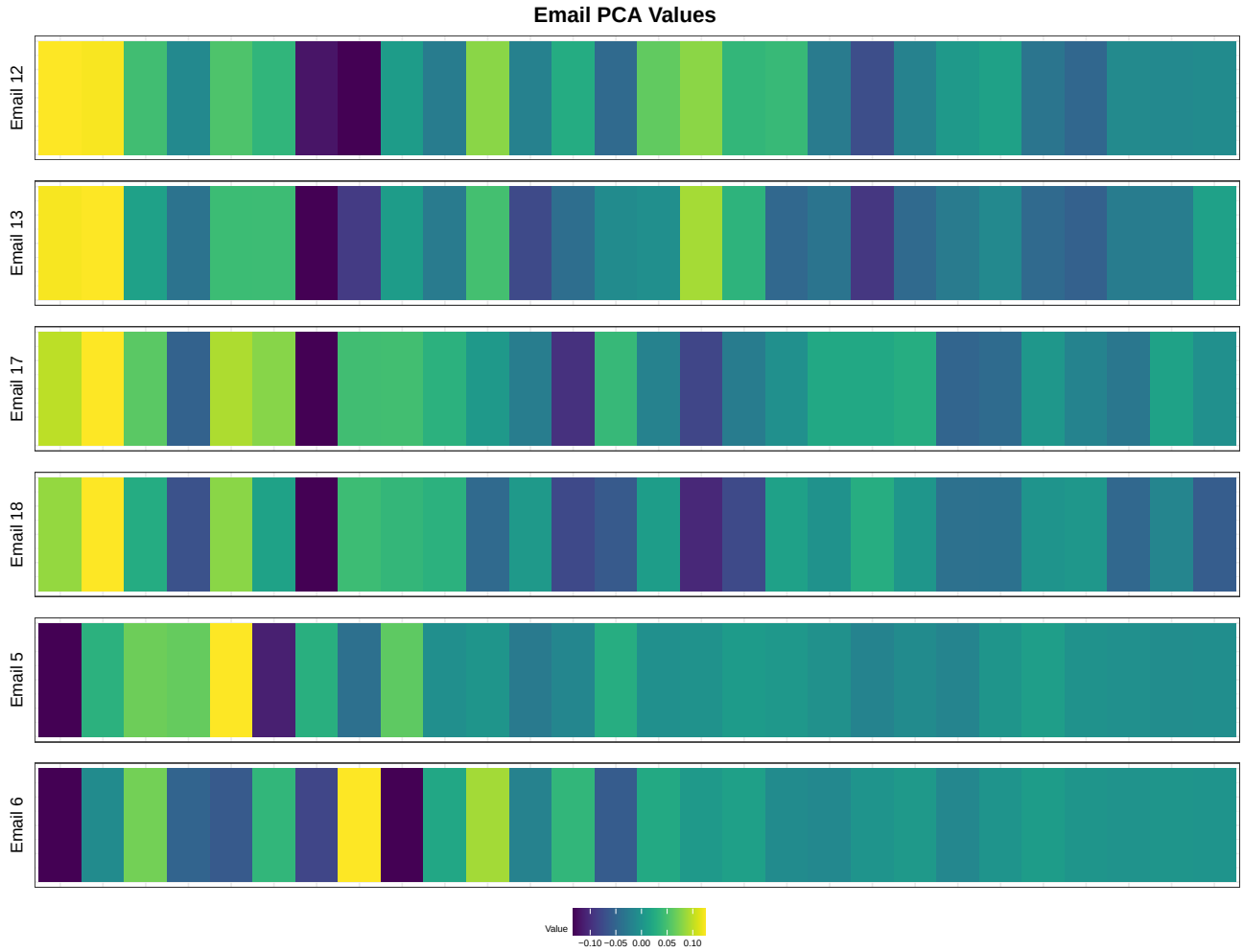


Figure 4: **PCA Representation of Six Example Emails.** This figure shows the result of PCA reduction of the Ada Embeddings on our 6 example emails. In order to capture as much variance in the contextual embeddings as possible, we use 28 variables from PCA reduction in our model. In this figure, Email 12 has subject line "Extra 20% off Clearance!" Email 13 has a subject line "Extra 30% off Clearance!" Email 17 is an "Extra 30% off Clearance - Online Only!" Email 18 is an "Extra 40% off Clearance - Online Only!" Email 5 is a "Happy Birthday!" email. Email 6 is a "Happy Memorial Day!" email. As shown, there is commonality between the emails that show the lowest Euclidean distance, such as Emails 12 and 13.

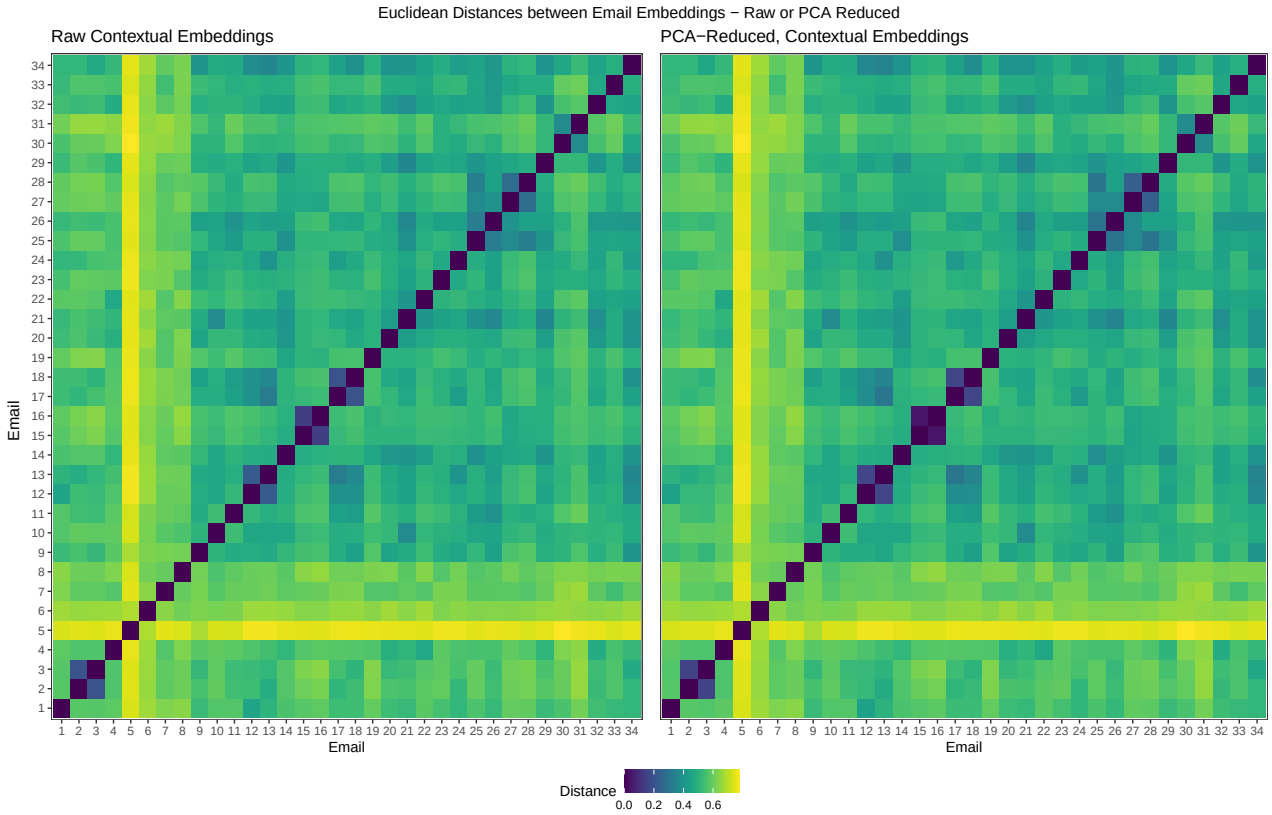


Figure 5: **Euclidean Distance Measure of Ada Embeddings for 34 Email Subject Lines.** This figure shows the Euclidean distances between the Ada Embeddings of our 34 email subject lines. The left panel shows the Euclidean distances based on the raw Ada embeddings. The right panel shows the Euclidean distances based on the PCA-Reduced, Ada embeddings. Darker blue regions are a distances between 0 and .2. Bright yellows correspond to a distance over .6. As shown the majority of our emails have distance measures between .4 and .2. The largest outlier is Email 5, the Happy Birthday! email, which has distance measures above .65. The figure shows that our emails show great saturation in the feature space, given the close proximity to each other. Most importantly, the use of PCA reduction maintains the natural ordering of distances found in the raw embeddings.

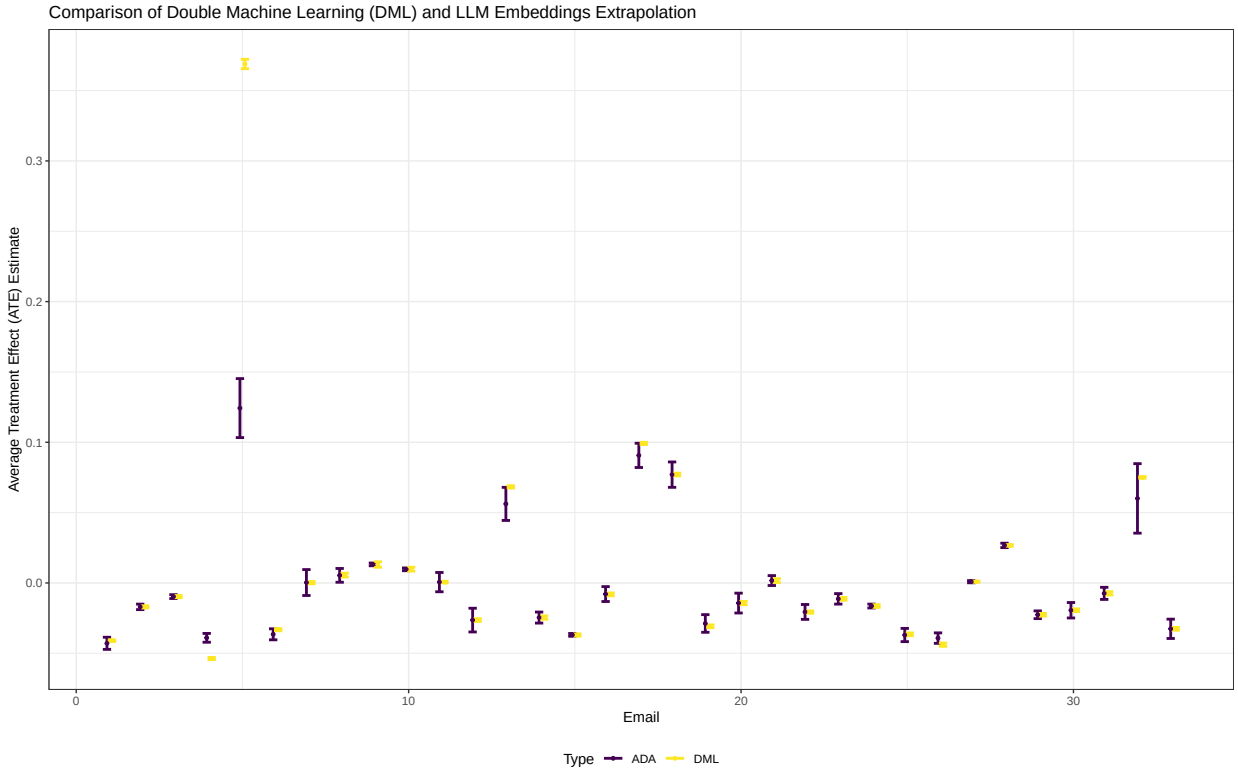


Figure 6: **Comparison of Average Treatment Effect (Ada - Extrapolated Score) using Ada Embeddings against the Doubly Robust (DML) Estimates of the Unconditional Purchase Amount.** This figure shows the performance of using Ada Embeddings to predict the average treatment effect (ATE) outcome of newly generated email creatives. We perform a leave-one-out exercise in our dataset, where each focal email is left out, our hold out email, and used as a new email creative. We train an optimally tuned XGBoost algorithm on the remaining 32 email doubly robust scores with the feature variables being the PCA measures of the remaining emails. Once estimated, the mean and standard error of the hold out email is calculated. For comparison, we use the ground truth, doubly robust score of the hold out email as a benchmark for comparison. Error bars are plotted at the 95% level. This figure shows that using the Ada Embeddings provides a viable set of features to predict new email creatives, as approximately 93% of our emails show overlapping 95% confidence intervals.

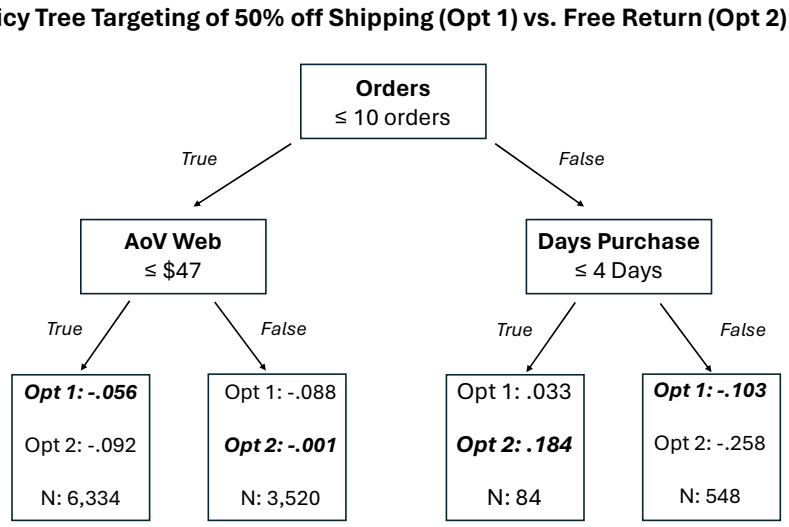


Figure 7: **Policy Tree Targeting Exercise of Heterogeneous Treatments.** This figure shows the decision tree result of the Policy Tree Algorithm (Athey and Wager, 2021) recommending targeting heuristics of two newly generated emails: *Limited Time Offer: 50% Off Shipping on All Products!* (Opt 1) and *Discover New Products: Free Returns on Every Order!* (Opt 2). Individual-level scores were calculated using an optimally tuned XGBoost Algorithm that contains both PCA-reduced Ada embeddings and pretreatment features shown in Table 2. We randomly sample approximately 10,000 individuals for the policy tree construction. Next, we calculate the new, individual-level scores for the two new emails generated by ChatGPT. Policy Tree then provides the heuristics of targeting based upon the calculated doubly-robust scores between the two different emails. Continuous variables were into 50 bins based upon the distribution of each variable. All pretreatment variables were used to construct the policy tree. As shown, there are clear targeting implications. We show the group average treatment effects (GATES) in the final nodes of the decision tree from policy tree as well as the number of observations in that node (N). We denote the chosen policy tree strategy in bold and italics within the final node. This figure shows the usefulness of targeting novel heterogeneous treatments based on pretreatment covariates. Lack of targeting results in a loss from the baseline email of -$717.80 for Opt 1 and -$739.86 for opt 2 versus a loss from the baseline email of -$425.77 from the targeting shown in this figure, which is a 40% improvement.

Web Appendix

A.1 Performance of Different Text Encoders

Keeping the above two requirements in mind, we now discuss some popular estimators of H_T, and provide guidance on which one is expected to be perform best. The earliest and most common practice is to estimate H_T by human codification. Here researchers pre-define a set of characteristics that describe the treatment object, typically drawing upon domain expertise. There are two key limitations of this approach. First, it is hard for human agents to fully encode the information in unstructured data, e.g., long text documents, video, audio, etc. In other words, this approach does not easily scale. Second, it is easy for human-beings to over-fit the encoded information based on their own experiences, such as researchers and customers having different interpretations of the key aspects of promotion emails.

Given the rapid progress in developing scalable ML algorithms, researchers and practitioners have gradually transitioned to algorithmic encoders over the past 15 years. These modern encoders offer several benefits over human codification. First, they are able to encode unstructured data at scale. Second, they can be trained with massive amounts of inputs (including the entire corpus of the internet, subject to computing and copyright constraints). Third, the fact that they are pre-trained on a global domain can lead to less concern for over-fitting when used in a particular context (such as novel treatment prediction).

There are now many algorithmic encoders to choose from. For instance, in the domain of text data, researchers can choose between topic modeling, n-gram, and embedding models (among other options). Topic modeling refers to an ML-based algorithm of clustering or grouping objects together based upon similarities in the text they include. By finding a latent set of topics, the algorithm then classifies the pieces of text into different groups. Alternative, n-grams refers to an algorithm that develops a set of features based upon a sequence of n adjacent words obtained via text classification.

To confirm our conjecture that the contextual embeddings can better model H_T than other candidate method, we consider four different approaches: human codification, topic modeling, bi-grams (n-grams for $n = 2$), and contextual (Ada) embeddings extracted from ChatGPT. To illustrate what the email subject lines look like in the contextual space after being encoded, Figures A.1 – A.3 plot the encoded contextual information of 6 example emails for human codification, topic modeling, and bi-grams, respectively.[27] In these figures, each vertical, colored line represents 1 of the contextual features that describes an email subject line. We can see that the coarsest information is generated by topic modeling. This is not surprising because we only have 34 emails, so there are not many topics to generate. On the other hand, Ada embeddings generate the finest information, implying that this approach is able to capture subtle differences in the contexts of these subject lines. For example, Email 12 states "Extra 20% off Clearance - Online Only!", while Email 13 is "Extra 30% off Clearance - Online Only!". Even at this granular level, we see similar patterns in these two emails across the embeddings: the pattern of light vs. dark is very similar between the two. Contrast that with Email 5, an email subject line that merely states "Happy Birthday!". While there are some similarities across the spectrum of embeddings, the relative positions of

[27] We include these for comparison with our Ada Embedding Figure shown in the main body of the paper.

light and dark bars are quite different between Email 5 and Emails 12 and 13. The contrast between these embeddings, while residing in the same space, provides some insight into how the fundamental characteristics of these messages are mapped into the embedding space.

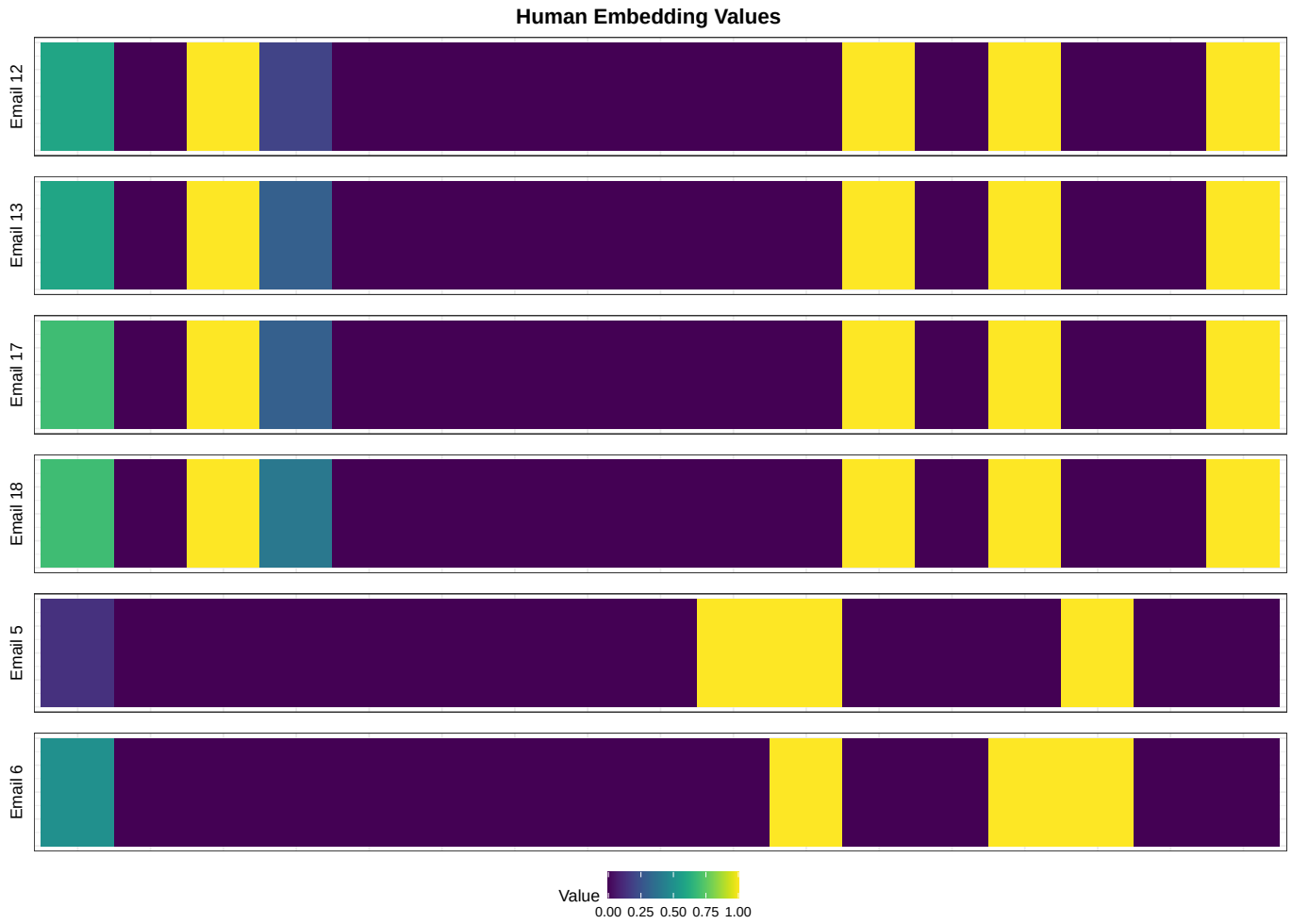


Figure A.1: **Human Codification Embeddings of 6 Example Emails.** This figure shows the Human Codification Embeddings of 6 example emails. In this figure, Email 12 is a Extra 20% off Clearance! email. Email 13 is a Extra 30% off Clearance! email. Email 17 is a Extra 30% off Clearance - Online Only! email. Email 18 is a Extra 40% off Clearance - Online Only! email. Email 5 is a Happy Birthday! email. Email 6 is a Happy Memorial Day! email.

We first check the performance of the Ada embeddings versus the other candidate methods in terms of in-sample fit (for this initial exercise, we do not leave out any emails). Figure A.4 presents the results: we bootstrap the RMSE values across 50 simulations for each method and present the percentage difference in RMSE of the other methods versus the Ada-PCA values. The methods are ordered from left to right in terms of their increasing percentage differential in prediction of the RMSE. In this exercise, we see that topic modeling performs the worst, which is not surprising since it generates the coarsest contextual representation. The Ada embeddings are shown to perform as well as, if not better than, the other methods in capturing the essence of the high-dimensional treatment objects in our

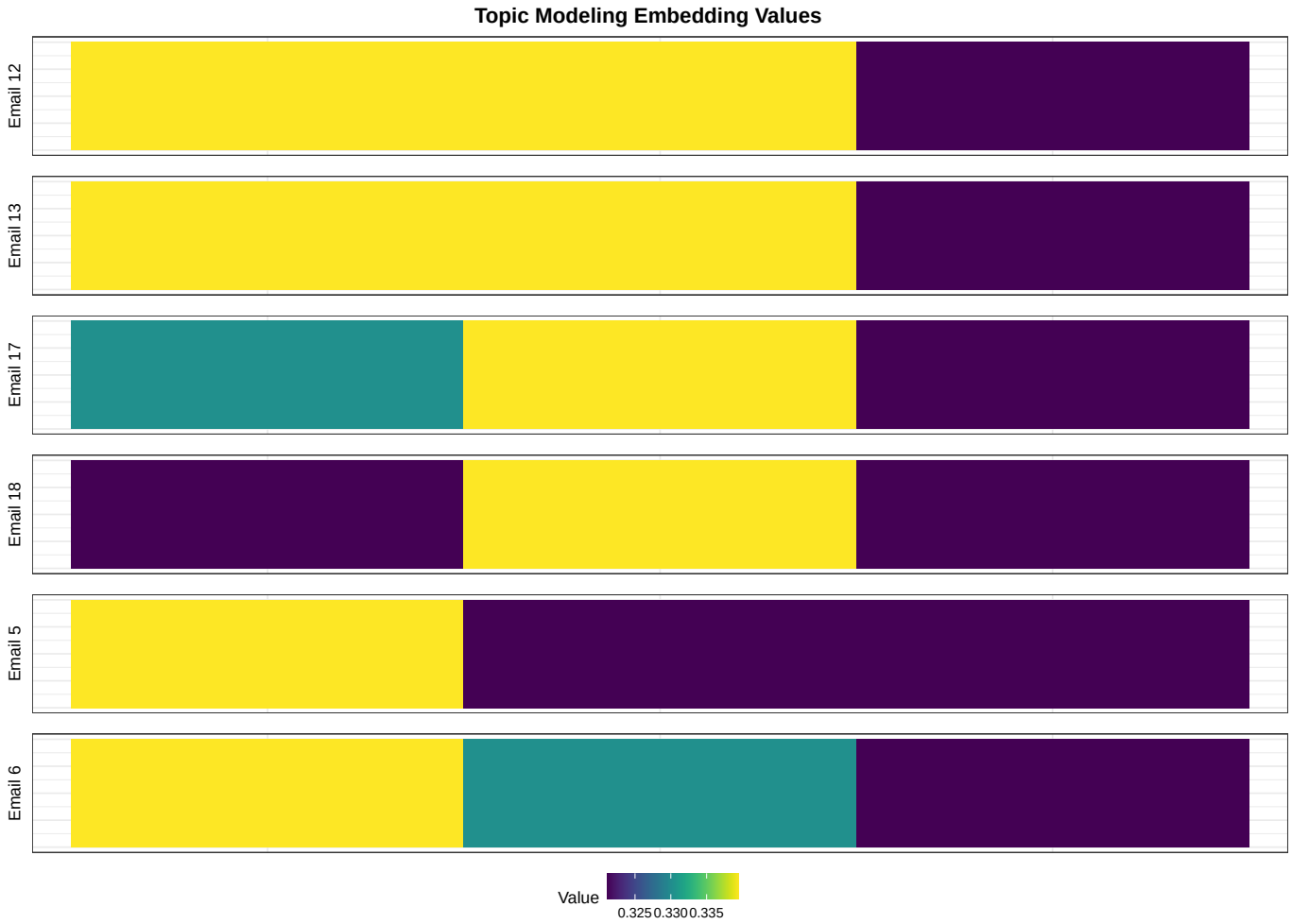


Figure A.2: **Topic Modeling Embeddings of 6 Example Emails.** This figure shows the Topic Modeling Embeddings of 6 example emails. In this figure, Email 12 is a Extra 20% off Clearance! email. Email 13 is a Extra 30% off Clearance! email. Email 17 is a Extra 30% off Clearance - Online Only! email. Email 18 is a Extra 40% off Clearance - Online Only! email. Email 5 is a Happy Birthday! email. Email 6 is a Happy Memorial Day! email.

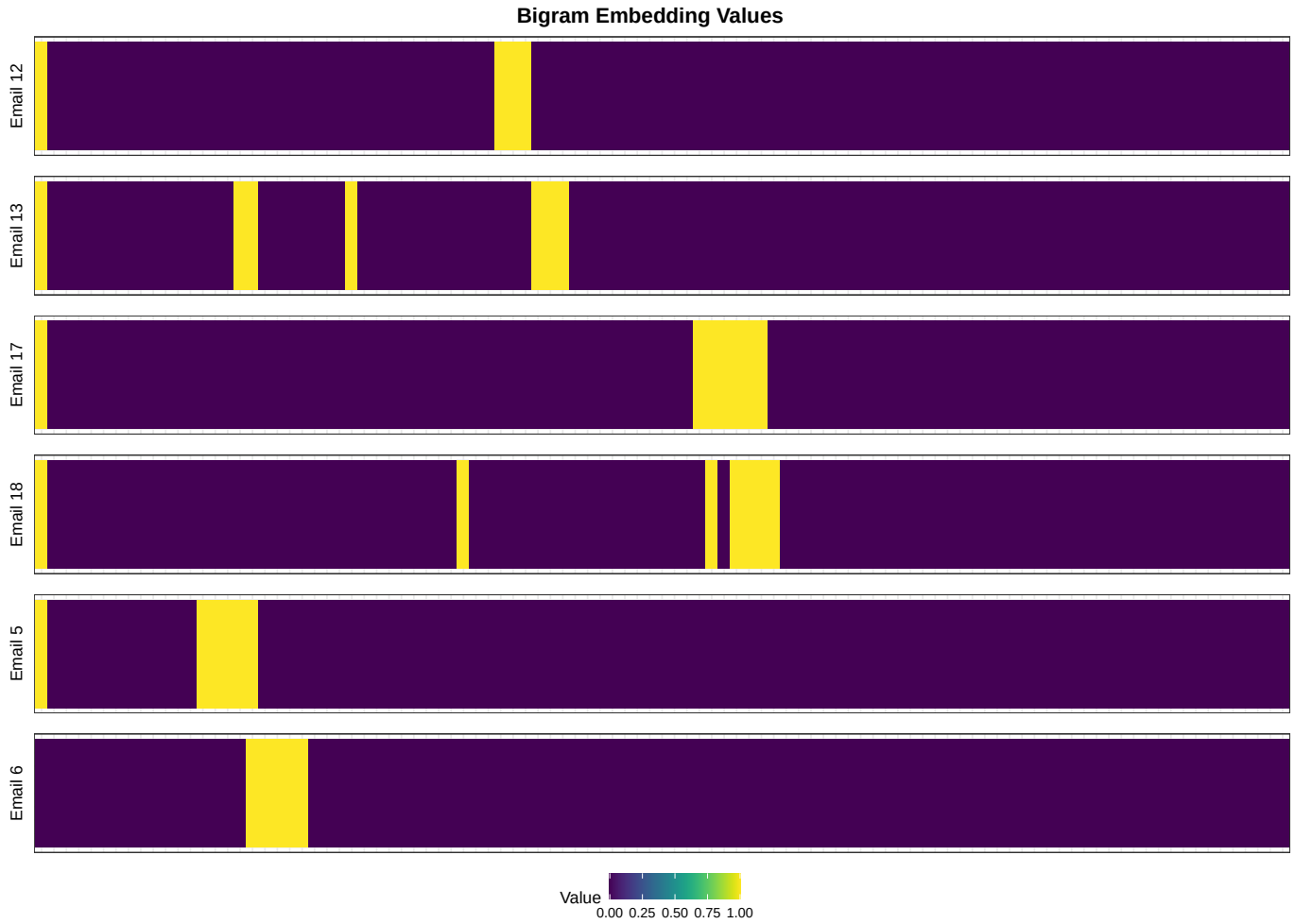


Figure A.3: **Bigram Embeddings of 6 Example Emails.** This figure shows the Bigram Embeddings of 6 example emails. In this figure, Email 12 is a Extra 20% off Clearance! email. Email 13 is a Extra 30% off Clearance! email. Email 17 is a Extra 30% off Clearance - Online Only! email. Email 18 is a Extra 40% off Clearance - Online Only! email. Email 5 is a Happy Birthday! email. Email 6 is a Happy Memorial Day! email.

setting. Though there is not much material difference between the fit of the four methods, our setting is an ideal one for the other methods to succeed. Email subject lines have a limited amount of text, and that text has limited variability. Other candidate methods can easily capitalize on this lack of variability to generate appropriate feature vectors. Even so, we find that the simple use of the Ada embeddings from ChatGPT provides the best solution in this case. Given the constrained set of text, we argue that in situations with a richer depth of text, the benefits of the contextual embeddings will only grow compared to other candidate methods.

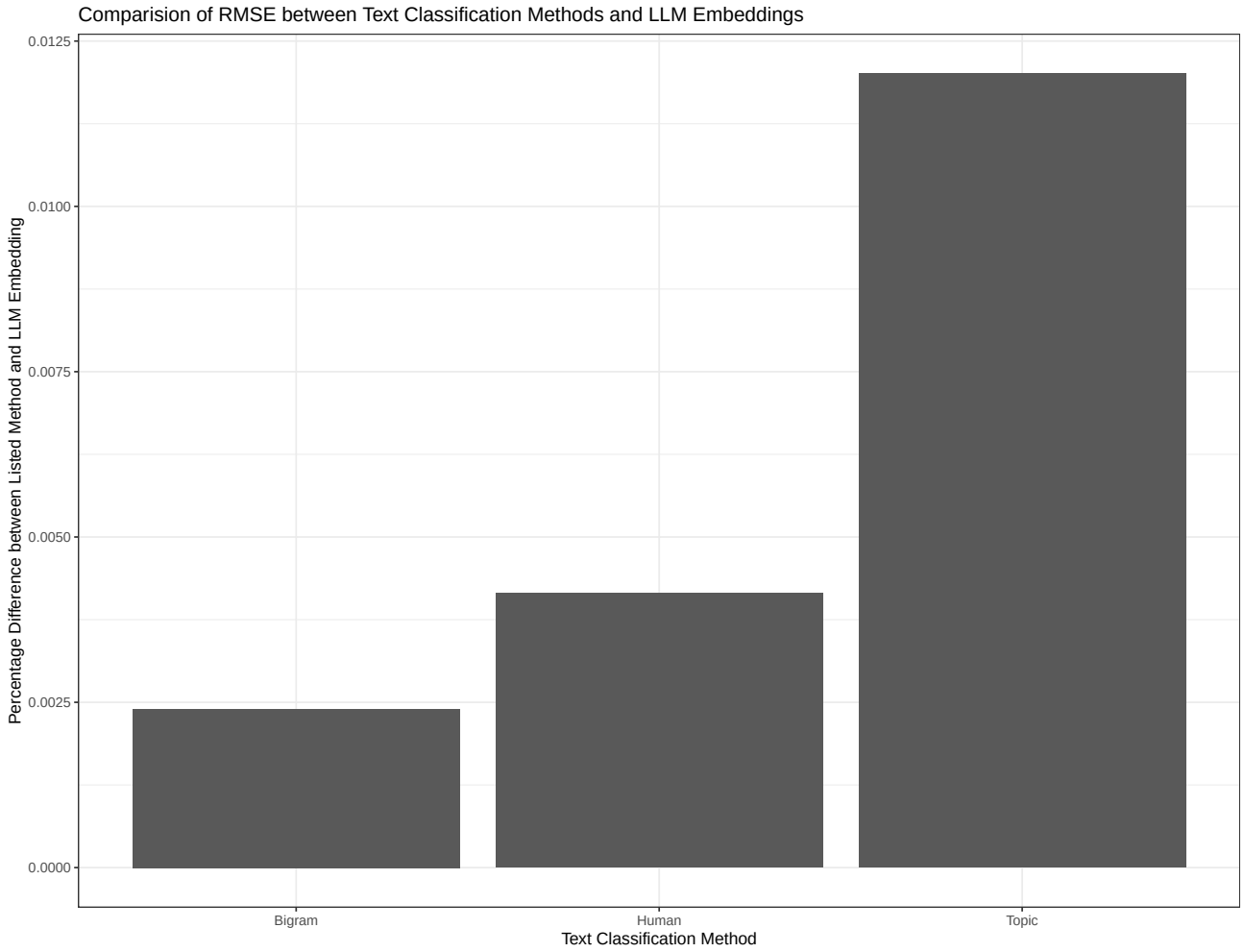


Figure A.4: **Comparison of Ada Embeddings as Feature Variables Against Commonly Used Text-Classification Methods.** This figure shows the relative performance of the RMSE from using Ada Embeddings versus three other candidate methods of text-classification. We use the bigram method, topic modeling, and human codification to generate the feature variables for an optimally tuned XGBoost Model. We bootstrap our sample 50 times and then calculate the mean RMSE performance for all four methods. The figure shows the percentage difference between the Ada embeddings versus each candidate method. In all three cases, the Ada embeddings perform better than the three candidate methods in having a lower RMSE.

A.2 Contextual Embedding Construction

For completeness, we include Figure A.5, which shows the cosine similarity between the email contextual embeddings. Though we focus on the use of Euclidean distance as our metric of feature saturation, this figure shows similar support with our definition using a different, commonly used metric.

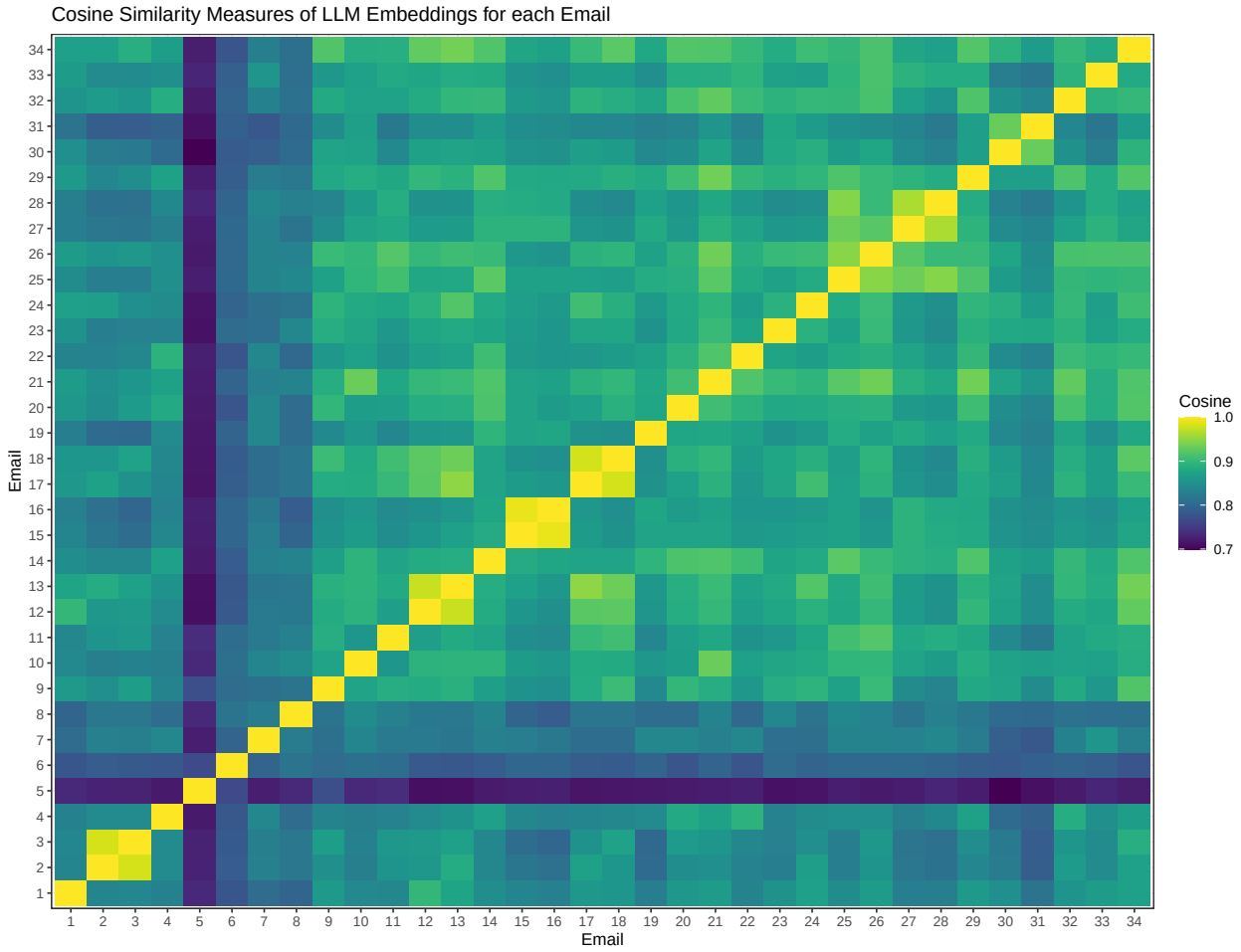


Figure A.5: **Cosine Similarity Measures of Ada Embeddings for 34 Email Subject Lines.** This figure shows the cosine similarity ratings of our 34 email subject lines. Darker blue regions are a cosine similarity measure of approximately .7. Bright yellows correspond to a cosine similarity measure near 1. As shown the majority of our emails have cosine similarity measures between .85 and 1. The largest outlier is Email 5, the Happy Birthday! email, which has cosine similarities around .75 or lower. The figure shows that our emails show great saturation in the feature space, given the close proximity to each other from the cosine similarity measure.

A.3 XGBoost Information and PCA on Embeddings Robustness

In our empirical analysis, we've transformed the 1536-dimensional Ada embeddings into a set of 28 PCA-reduced embeddings. Ada embeddings lack direct human interpretability, so reducing their dimensionality doesn't result in any loss of crucial information for managerial insights, provided the reduced embeddings capture the majority of the variance. This reduction also facilitates the utilization of more parsimonious and scalable models on the transformed vectors. To determine the optimal number of PCA components, we systematically increased the number of components and examined their ability to explain variance. We settled on using PCA with 28 components, as it accounted for 99% of the variance, as illustrated in Figure A.6.

In our XGBoost model, we use a histogram-based algorithm to build the trees. Further, we leverage several key parameters: 'subsample', 'max_depth', 'colsample_bytree', and 'learning_rate'. Each of these parameters plays a crucial role in shaping the behavior and performance of our model. We meticulously tune these parameters to optimize the performance of our XGBoost model. Below, we detail the significance of each parameter and the range we explore during our experimentation.

- 'subsample': This parameter controls the fraction of samples used for training each tree. It ranges from 0 to 1.0 and represents the proportion of the dataset to sample randomly without replacement. A value of 1.0 means using all samples for each tree. We vary *subsample* from 0.5 to 1.0 in increments of 0.1.

- 'max_depth': This parameter sets the maximum depth of each tree in the ensemble. It controls the complexity of the trees and helps prevent overfitting. Larger values can lead to more complex models, potentially overfitting the training data. We vary *max_depth* from 3 to 8 in increments of 1.

- 'colsample_bytree': This parameter specifies the fraction of features (columns) to consider when building each tree. It ranges from 0 to 1.0 and represents the proportion of features to randomly select for each tree. Smaller values can help prevent overfitting by promoting diversity among trees. We vary *colsample_bytree* from 0.3 to 0.9 in increments of 0.1.

- 'learning_rate': This parameter controls the step size at each iteration while moving toward a minimum of the loss function. It ranges from 0 to 1.0. Lower values require more iterations but can result in better convergence, while higher values may converge faster but risk overshooting the minimum. We vary *learning_rate* from 0.1 to 0.6 in increments of 0.1.

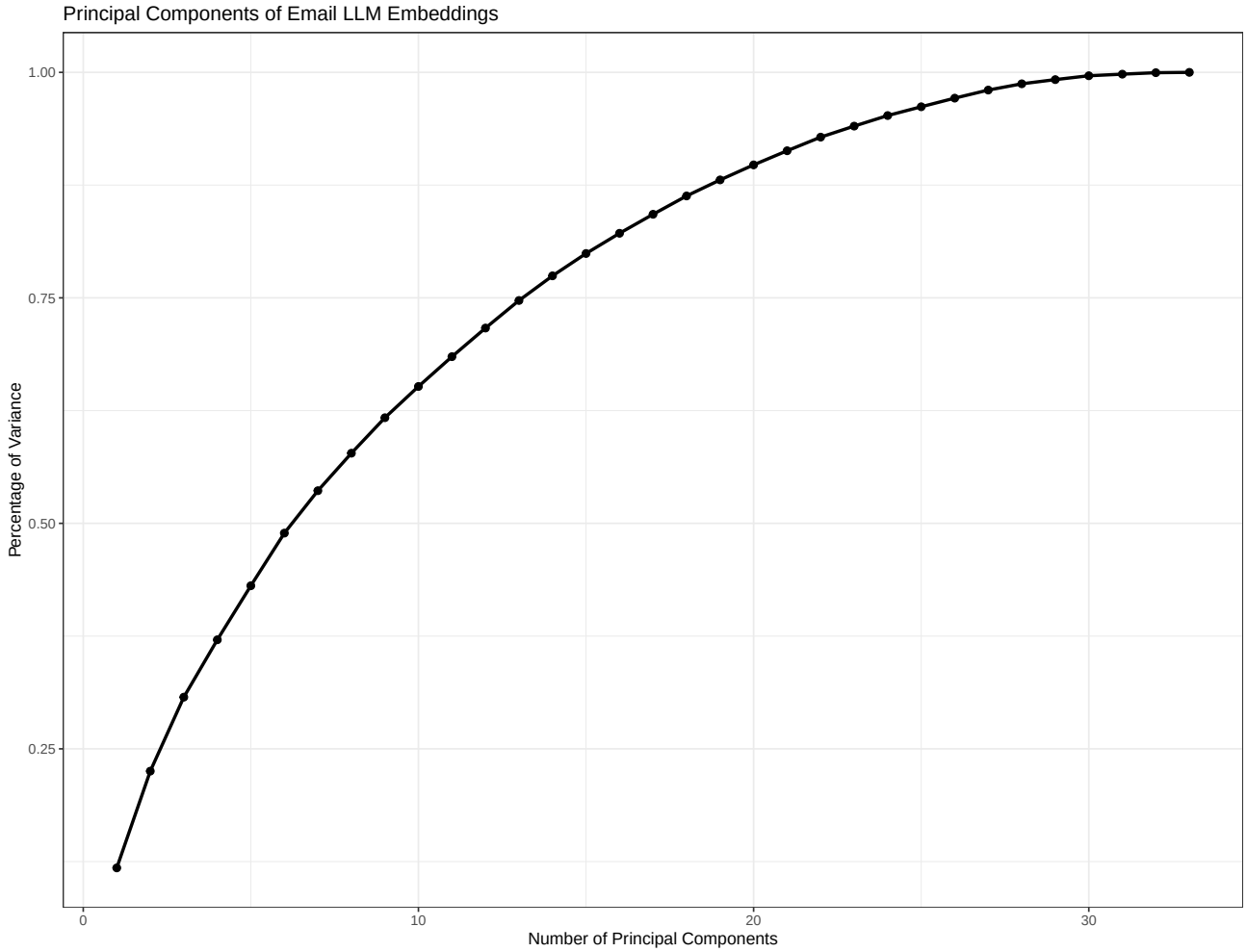


Figure A.6: **PCA Representation of 6 Example Emails.** This figure shows incremental gains of adding another component to PCA dimension reduction of our obtained Ada Embeddings for 34 emails. Our objective is to capture as much variance as possible for XGBoost to use in variable selection. From this plot, 28 variables accounts for approximately 99% of the variance, which is why it is selected. In addition, since the Ada embeddings are normalized, we show the decay in variance reduction is almost linear rather than geometric as normally shown in PCA analysis.

A.4 ATE Prediction with Demographics

For completeness, Figure A.7 presents the ATE_D versus the extrapolated ATE_{Exp} when we include customer demographics. Noticeably, the error bars for the prediction are larger when customer demographics are considered, which is a direct cause of lack of support in some regions of customer characteristics in the dataset. While this may be a cause for concern, our analysis of the GATES provide evidence that we are able to effectively recover the trajectory of heterogeneous scores of novel heterogeneous treatment.

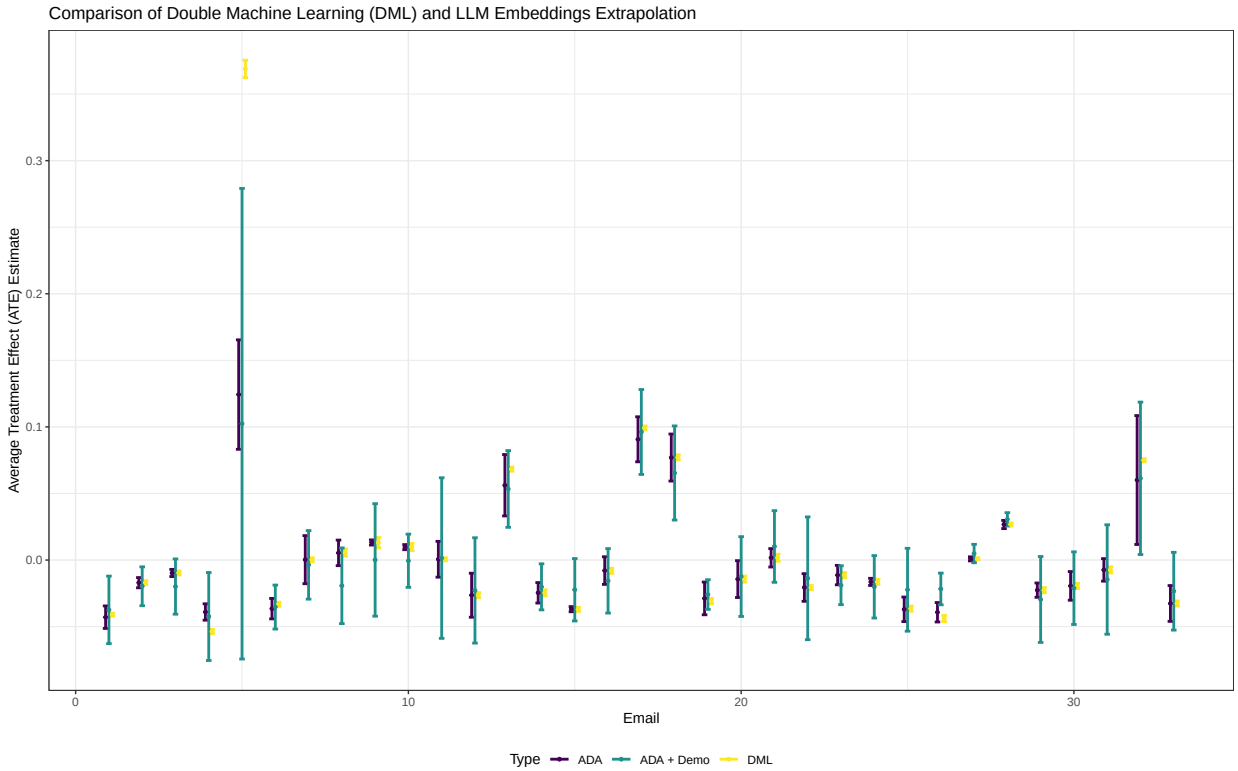


Figure A.7: **Comparison of Average Treatment Effect (Ada - Extrapolated Score) using Ada Embeddings against the Doubly Robust (DML) Estimates of the Unconditional Purchase Amount.** This figure shows the performance of using Ada Embeddings, with the inclusion of customer demographics, to predict the average treatment effect (ATE) outcome of newly generated email creatives. We perform a leave-one-out exercise in our sample, where each focal email is left out, our hold out email, and used as a new email creative. We train an optimally tuned XGBoost algorithm on the remaining 32 email doubly robust scores with the feature variables being the PCA measures of the remaining emails. The difference between this plot and the prior is the inclusion of customer-level characteristics in development of the extrapolated scores. Once estimated, the mean and standard error of the hold out email is calculated. For comparison, we use the ground truth, doubly robust score of the hold out email as a benchmark for comparison. Error bars are plotted at the 95% level. Noticeably, the error bars for the prediction are larger when customer demographics are considered, which is a direct cause of lack of support in some regions of customer characteristics in the dataset. While this may be a cause for concern, our analysis of the GATES (described in Section 4.5 provide evidence that we are able to effectively recover the trajectory of heterogeneous scores of novel heterogeneous treatment.

A.5 ChatGPT Prompting

For completeness, we are including the direct text prompt given to ChatGPT to generate our new emails. We used ChatGPT 3.5 for this analysis. After submitting the text prompt, we then asked ChatGPT to rank order the new emails based on its belief of their performance.

- Create subject lines of six email promotions where each can be about:

  Product

  Clearance

  Other Kinds

- In each of these email subject lines, use no more than three of the following components:

  May choose a percentage discount between 10% to 50%

  Cash Discount

  BOGO

  Free Shipping

  50% Free Shipping

  Free Returns

  Happy Birthday

  Mystery

  Code that should be used at checkout

- The following words can be used for embellishment:

  Extra

  Exclusive

  Exclaim

  Sale

- The length of each subject line should not exceed 60 characters with an average length of 50 characters.

- Do not use emojis.

New Creatives

Based on the above prompts, ChatGPT created the following new email subject lines:

- Exclusive Product Sale: Extra 20% Off Today!
- Clearance Event: BOGO Deals + 30% Off Products!
- Other Kinds Extravaganza: Code for Extra Savings!

- Limited Time Offer: 50% Off Shipping on All Products!
- Happy Birthday Special: Mystery Discount Awaits You!
- Discover New Products: Free Returns on Every Order!

A.6 ChatGPT ranking of Training Emails

To test the robustness of our findings that GPT is unable to properly rank email creatives, we use our doubly robust purchase amount orthogonal scores of our 33 emails to serve as the ground truth of email performance. Next, we ask GPT to rank the 33 emails based on their purchase performance, where 1 is the best and 33 is the worst. We simulate this exercise 30 times and perturb the ordering of the emails within the prompt to ensure there is no bias in the prompt engineering. Figure A.8 shows the results of this exercise. We order the emails from the best to worst performing based on our DML estimates on the Y axis. The X axis shows the difference between the Estimated Purchase Amount Rank, from the DML estimates, against the rank from our simulated GPT prompt. As shown, GPT is unable to properly rank the creatives based on performance. In fact, those emails which perform the best, GPT suggests that they would be worst and vice versa. This is congruent with the exercise from the main paper that examines the predictions of novel heterogeneous treatments derived by GPT.

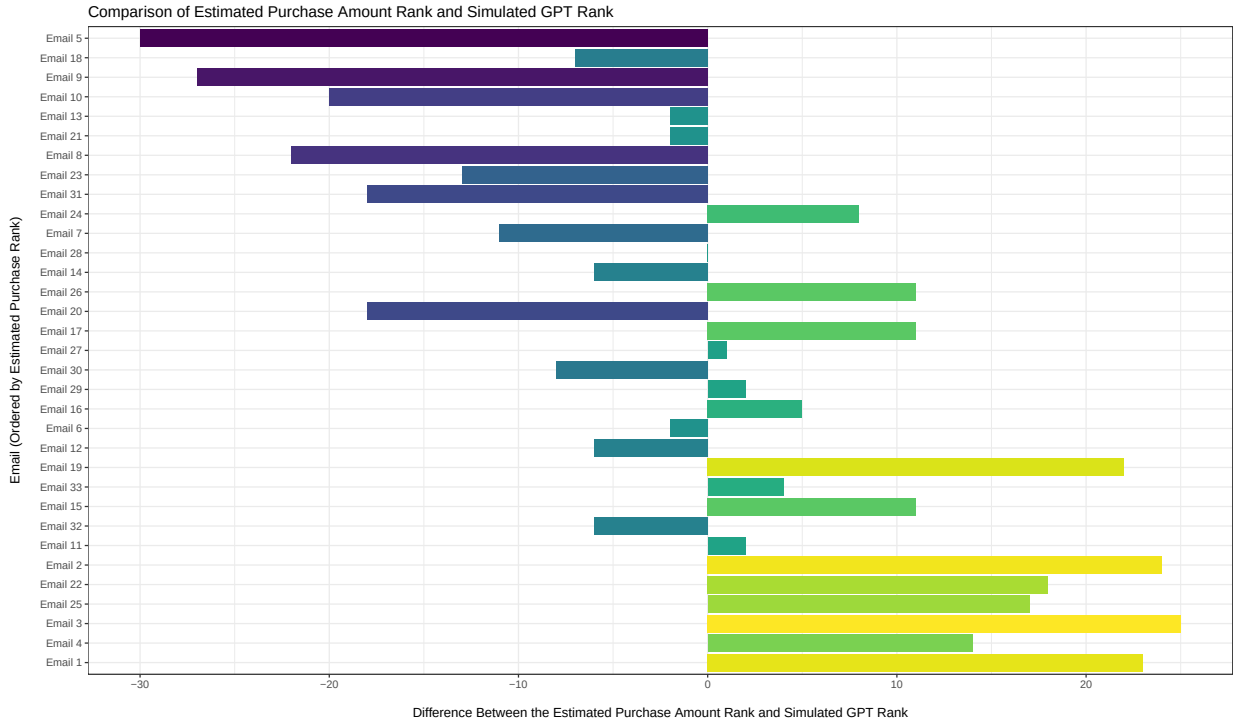


Figure A.8: **Differential between Estimated Purchase Amount Rank from DML versus Estimated Purchase Amount Rank from GPT Prompting.** This figure shows the difference in ranking suggested by our doubly robust orthogonal scores versus those suggested by GPT. Simulated GPT ranks are computed across 30 simulations where the final stated rank is the order based on the simulated rankings. We show the difference between the DML Ranking versus the GPT Ranking in the figure, where emails are ordered from best performing (top of the Y axis) to worst performing (bottom of the Y axis). As shown, GPT is unable to properly rank the best from worst emails.