Working Paper 23-062

Using GPT for Market Research

James Brand
Ayelet Israeli
Donald Ngwe

Harvard Business School

Using GPT for Market Research

James Brand
Microsoft

Ayelet Israeli
Harvard Business School

Donald Ngwe
Microsoft

Working Paper 23-062

The authors are grateful to Meng Yang for excellent research assistance.

Funding for this research was provided in part by Harvard Business School.

Using GPT for Market Research*

James Brand[†] Ayelet Israeli[‡] Donald Ngwe[†]

July 7, 2023

Abstract

Large language models (LLMs) have quickly become popular as labor-augmenting tools for programming, writing, and many other processes that benefit from quick text generation. In this paper we explore the uses and benefits of LLMs for researchers and practitioners who aim to understand consumer preferences. We focus on the distributional nature of LLM responses, and query the Generative Pre-trained Transformer 3.5 (GPT-3.5) model to generate hundreds of survey responses to each prompt. We offer two sets of results to illustrate our approach and assess it. First, we show that GPT-3.5, a widely-used LLM, responds to sets of survey questions in ways that are consistent with economic theory and well-documented patterns of consumer behavior, including downward-sloping demand curves and state dependence. Second, we show that estimates of willingness-to-pay for products and features generated by GPT-3.5 are of realistic magnitudes and match estimates from a recent study that elicited preferences from human consumers. We also offer preliminary guidelines for how best to query information from GPT-3.5 for marketing purposes and discuss potential limitations.

*The authors are grateful to Meng Yang for excellent research assistance.
[†]Office of the Chief Economist, Microsoft; jamesbrand@microsoft.com and donald.ngwe@microsoft.com
[‡]Harvard Business School; aisraeli@hbs.edu

1 Introduction

Large language models (LLMs) are a type of artificial intelligence designed to understand and generate human-like language. These models are trained on vast amounts of text data, which allows them to learn the patterns and structures of natural language. Large language models have a wide range of applications, from language translation and speech recognition to content generation and text classification. They are becoming increasingly popular in industries such as finance, healthcare, and marketing, as they are able to process and analyze large amounts of text data quickly. LLMs power several well-known AI-augmented solutions for coding (e.g., Github Copilot) and search (e.g., Bing, Bard), and a small number of studies have shown that they can also replicate limited real-world behavior, including voting (Argyle et al., 2022) and some economic experiments (Horton, 2023).

In this paper, we investigate how LLMs (in our case, Generative Pre-trained Transformer 3.5, "GPT-3.5" or "GPT" henceforth) can be used as a tool for market research.[1] GPT's training data includes information from numerous sources on the internet, which may include product reviews, messaging boards, and other online forums with contributions from a wide range of consumers discussing the products they shop for and purchase. Because GPT and similar LLMs are trained to respond to queries and prompts with the most likely next sequence of text, we expect that the responses GPT provides to market research surveys will reflect the types of responses that would have arisen among the customers in the training data. Together, these components suggest that GPT may be an invaluable source of insight into consumer preferences due to its ability to mimic or replicate human responses.

Existing tools for market research, such as conjoint studies, focus groups, and proprietary data sets can be expensive. If LLMs can generate responses that are consistent with existing studies on human subjects, then they may also be able to serve as a fast and low-cost method of providing the information typically generated by conjoint studies and other customer surveys. As major tech companies have begun to combine LLMs with tools for searching and synthesizing information from the web, one might imagine using LLMs to develop marketing or pricing strategies prior to the launch of a new product, and then iteratively querying LLMs over time to evaluate product-market fit and modify the marketing strategy. In a way, consumers are surveyed indirectly — through their part in forming the text on which LLMs are trained.

We emphasize that, *ex ante*, it is unclear what we should expect to learn from GPT's responses

[1]Note that we access GPT-3.5 directly using OpenAI's API rather than through the more widely-used ChatGPT interface. ChatGPT is an application that uses a variant of GPT which has been optimized for dialogue and following user instructions rather than the type of text completion we focus on here. See Appendix A for code and details of our implementation.

to typical consumer survey questions. Product reviews, for example, which are likely present in the training set for GPT, may reveal something about customers' stated preferences for products but may not always mention prices or other key attributes of the product or of the decision-maker (e.g, income or demographics). When GPT is offered a $100 candy bar, will it know to decline? When it is offered a choice between a $1 plain vanilla bar and a $2 chocolate fudge bar, will it know how to make the trade-off? Moreover, even if GPT can generate reasonable responses to each isolated question, will its responses *across* different questions be internally consistent in the ways we expect consumers to be? Evaluating these issues is key to understanding the potential value of GPT and other LLMs for almost any marketing analysis, and is the focus of this paper.

A priori, it is also unclear whether GPT's training set can generate useful responses. A large literature documents the differences between customer surveys, which elicit stated preferences over bundles of goods, and real-world demand data, in which customer preferences are revealed by their actual choices. (See, for example, Kroes and Sheldon, 1988 and Johnston et al., 2017.) GPT's training set contains aspects of both: consumers comment online about actual or prospective purchases. However, posted comments about purchases are neither a representative sample of actual sales data nor prompted by typical consumer survey questions. This aspect of the training set, together with the opacity with which GPT forms responses to prompts, motivates our investigation into the usefulness of GPT for market research.

Our findings are encouraging. We begin by measuring the extent to which GPT exhibits fundamental, well-established, properties of consumer demand by conducting four studies. In each study, we provide GPT with a series of prompts, varying key features of the choice setting or of customer attributes (e.g., prices, prior purchases, income). The results suggest that the preferences implied by GPT's responses are consistent with downward-sloping demand curves, diminishing marginal utility of wealth, and state dependence. Next, we explore the realism of GPT's responses, first by directly eliciting willingness-to-pay (WTP) for products in multiple categories and then by estimating WTP for product attributes via three approaches. Overall, the resulting WTP estimates seem realistic, both in magnitudes and in distribution. In particular, we show that a conjoint-like approach to preference estimation yields results that are strikingly similar to those found in a recent survey of real consumers conducted by Fong et al. (2023). Together, our results suggest that GPT potentially provides an alternative means for marketers to learn about consumer preferences in a fast, low-cost, and iterative manner. Whereas a survey of real customers may cost many thousands of dollars and take weeks or months to implement, each of our studies ran in a matter of minutes or hours, and the total cost to generate all the data in the paper was under $100.

Although these results are promising, research in this area is preliminary, and more work is needed to identify best practices for learning customer preferences from LLMs. In Section 4, we provide some guidance on the limitations of the approach and issues we encountered while conducting our studies. For example, GPT is sensitive to the phrasing of prompts, and while many of the behaviors we show here are robust in direction, their magnitude can differ depending on the precise prompt we provide. We found that while asking GPT for a “single price in dollars" generated responses in whole dollar amounts, “a single price in dollars and cents" or just “a single price" solved this problem. We also found that, like human survey participants, GPT exhibits response order bias and is much more likely to choose the first option in a binary choice than the second. Thus, although we find success from a GPT-based conjoint (Section 3.2.2) with minimal prompt engineering and no fine-tuning, we advise other researchers to validate our findings in their own contexts before relying on GPT surveys alone for estimates of consumer preferences.

1.1 Existing Literature

A nascent but growing literature studies the economic benefits of LLMs from multiple angles. Most relevant to our study is Horton (2023), which demonstrates that various OpenAI LLMs provide responses to economic scenarios in ways that are consistent with intuition and experience. Horton makes the distinction between stated and revealed preferences and concludes that the corpus on which LLMs are trained is more likely comparable to revealed preferences, focusing on classic experiments from behavioral economics. He also compares GPT to a random number generator, which is related to our approach. We focus on the distribution of prompt responses rather than a single draw.

Prior work has identified specific means by which machine learning (ML) and generative AI models can benefit marketing practice. Conceptually, the paper closest to ours is Netzer et al. (2012) who extract customer preferences from text. Timoshenko and Hauser (2019), and Burnap et al. (forthcoming) demonstrate how marketing managers can use ML/AI approaches to improve the efficiency of intensive, manual, and costly processes. In the context of generative AI, Li et al. (2023) demonstrate how to use LLMs to construct perceptual maps by querying GPT about brand similarities, querying GPT one to five times for each pair of brands and comparing its responses to those of humans. We contribute to this stream of the literature by further demonstrating how a widely available generative AI tool can extract consumer preferences.

The broader literature on generative AI has identified several means by which AI can improve productivity. Peng et al. (2023) and Noy and Zhang (2023) use experiments that show that access to generative AI allows participants to complete tasks faster and with higher quality than those without access to the AI. Mollick and Mollick (2023) show how GPT can be used to improve teaching effectiveness. Our work similarly has an eye toward increasing productivity, albeit at the level of marketing methods rather than the individual user level.

2 Research Design

2.1 GPT-3.5

In this paper we focus on GPT as a cutting-edge example of the broader LLM technology. GPT was developed by OpenAI and released publicly in 2020, and OpenAI maintains a public API that makes it easy to submit many prompts quickly from Python or Julia[2] and to receive many different responses at once for each prompt. One key difference between our study and common illustrations of LLMs' capabilities to date is our focus on the distributional nature of LLM responses. Workers use an LLM to accelerate or improve their own output because of its ability to reliably provide a valuable response quickly. The process for querying LLMs in these contexts tends to consist of either autocomplete-style responses, where the LLM provides only a single response to the worker, or a conversational or interactive environment where the worker might purposefully submit similar queries a few times in a row to explore different alternatives. However, this form of interaction with LLMs is not ideal for understanding customer preferences, which is the focus of our work here.

2.2 GPT as a Simulator of Human Responses

Language models like GPT have been trained to predict text that would be written by humans, mostly on the internet, in response to a "prompt" which provides contextual information (Ouyang et al., 2022). After training, these models are then "aligned" to serve as more effective, truthful, and ethical assistants for purposes like developing chat-based assistants. Our focus is on these text completion models *prior to alignment*. Our hypothesis is that, when we induce GPT to provide comparisons between products in a simulated market research study, the responses it provides reflect the learned distribution of responses from the consumers that compose its training data. This is independent of GPT's ability to produce factual information on request, and relies solely on the assumption that a model that can accurately predict how humans respond to sufficiently many contexts must also reveal some of the preferences of the

[2]We used GPT-3.5 model "text-davinci-003" for our studies, which ran in March 2023, and accessed the API using Julia. See Appendix A for a code sample.

humans it aims to represent.[3] Hence, our approach to performing market research using GPT is to treat the model as a synthetic consumer rather than a source of knowledge.

Other recent work has also demonstrated early success using GPT as a tool to simulate human responses. Horton (2023) conducts economic experiments using GPT-3, Argyle et al. (2022) simulate samples of political preferences, and Aher et al. (2022) simulate psychological studies including the well-known Milgrom shock experiment. These studies focus on comparing the distributions of simulated responses to those from humans and in general find encouraging similarities between the two. While this type of comparison is also an interest of this paper, we emphasize that some of our results demonstrate a deeper, emergent, level of human simulation. Market research concerns not only what customers will say about their preferences, but what their choices reveal about those underlying preferences when economic models are estimated using the resulting data. Analysis of the latter involves subjecting humans to multiple questions with different contexts, and requires humans to behave in ways that are, for the most part, internally consistent. Thus, while we expect that GPT's responses to marketing questions will be qualitatively similar to humans, key questions still remain as to whether GPT's responses to these types of market research surveys will provide estimates of preferences that are realistic and consistent with estimates using human-generated data.

2.3 Querying GPT for Market Research

As with much of the empirical marketing and industrial organization literature, we wish to study the impact of changes in the attributes of goods on choice probabilities and market shares, which normally requires data from many randomly sampled customers or markets. This means that for each set of goods we consider, we need to query GPT hundreds of times, and our goal is for GPT to generate a distribution of responses rather than repeat a single one. To this end, we set the "temperature" on GPT to its maximum value for text completion (1.0)[4] for all studies in an attempt to maximize variation across responses.[5] Our prompting approach then proceeds as follows. In each study, we prompt GPT to fill in the responses of a survey question as if it were a customer shopping in a category of interest who was randomly selected to participate in a survey. We describe any relevant attributes of the customer (e.g., annual income), offer one or two products for this customer to consider purchasing, and then remind the customer

[3]"Hallucination" is the term used to describe cases where LLMs produces incorrect information, which is often of interest when using LLMs via chat-based interfaces or LLM-augmented search. Because we are not querying GPT for facts, we do not consider hallucination to be of critical importance for our research question.

[4]For chat completion, the maximum is 2.0, which generates "creative" responses.

[5]We echo Horton (2023)'s observation that " 'natural' human variation in preferences does not exist in LLMs unless they are endowed with differences." How setting the temperature in LLMs and thus increasing stochasticity relates to random sampling of human subjects is an interesting question that we leave for future work.

that they can always choose not to make a purchase. We then ask GPT to fill in the response of this chosen customer by ending our prompt with "Customer:". We submit each of these prompts to GPT hundreds of times and aggregate the responses to construct our measures of interest.[6] In crafting the prompts for our querying approach, our main goal was to demonstrate the usability of GPT as a market research tool, rather than fine-tuning the prompts. We discuss key learnings about GPT prompts in Section 4, and leave prompt engineering and optimization for future work.

2.4 Data Parsing

After collecting GPT's responses, we parse and aggregate them to analyze the results. We aimed to craft prompts that result in responses that are concise and contain only the information we require (see Section 4 for discussion). However, GPT at times tends to be verbose and provide lengthier comments, which require some coding to parse the output into a useful format (for example, even when asking for a single maximum price, GPT may respond: "15, any more than that and I wouldn't be getting a good deal for my money"). Appendix Section C presents examples of GPT responses.

3 Results

In our first set of studies, we study whether GPT's responses broadly align with predictions from economic theory. In our second set of studies, we compare GPT's responses to benchmarks that are representative of established market research tools: WTP measurement and conjoint analysis.

3.1 Testing Predictions from Economic Theory

We investigate four fundamental properties of consumer demand that are both prevalent in economic theory and widely documented in the economics and marketing literatures. We show that, by and large, GPT's responses align with expected (and observed) consumer behavior. For each of these studies, we prompt GPT as described in the previous section, varying attributes of the offered choice and aggregating hundreds of responses for each query.[7] The exact prompts we use and further details are provided in Appendix Section B.

[6]We include examples of the prompts we used in Appendix Section B.

[7]In the choice queries, we query GPT 300 times for each product and price level.

3.1.1 Study 1: Downward-sloping demand curve

A fundamental feature of economic models is that price elasticities for typical goods are negative and demand curves are downward-sloping. Given the importance of this feature in most empirical and theoretical work in economics, we begin our studies by establishing how GPT responds to price changes for a single good, holding everything else constant.

Figure 1 presents the results of three separate exercises to explore the shape of the demand curve from GPT responses.[8] First, in Figure 1a, we offer the GPT customer a binary choice between a single laptop (Surface Laptop 3) and the no-purchase option, varying the price of the laptop from \$749 to \$1,249.[9] In this simple scenario we find that the demand curve is trending downward. When the price of the laptop is below \$1,000, the fraction of customers choosing to purchase the laptop is nearly 10% larger than when the price of the laptop is above \$1,000. However, the decrease in demand is quite small and seemingly unrepresentative of typical human consumers.

We note that this and the succeeding demand curves implied by GPT are generally not *monotonically* downward-sloping. We interpret this characteristic as a feature and not a bug: much work in economics and marketing has documented several possible relevant factors, including context effects, preferences for round numbers, prices as a signal of quality, and left-digit effects (e.g., Payne (1982), Thomas and Morwitz (2005), Schindler and Kirby (1997), and Gerstner (1985)). The shape of demand that we find may be influenced by these and possibly as yet undiscovered phenomena, in addition to more familiar predictions from economic theory.

Our second and third exercises focus on GPT's choice among multiple options. In Figure 1b, we offer GPT an alternative laptop (selected to be a close substitute of the Surface 3) at a fixed price of \$999 while varying the price of the Surface 3 laptop.[10] Here we find a much steeper demand curve. Notably, although the demand curve is downward sloping throughout much of the price range, it exhibits a particularly sharp drop in demand for the Surface option at around the price of the reference good.

In Figure 1c we ask GPT to choose between two toothpaste brands (Colgate and Crest), varying prices between \$2 and \$6, with the reference good's price fixed at \$4. We note two takeaways

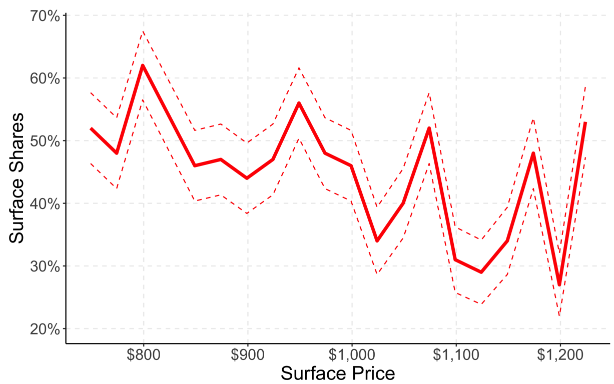
[8] Although GPT represents unique challenges for characterizing sampling variance, in all tables and figures herein we calculate standard errors as if our data were generated by randomly sampled consumers. We view this as a useful baseline, though we note that future work may wish to explore alternative approaches to inference in these settings.

[9] A sample of responses that GPT provided to this prompt appears in Appendix Section C.

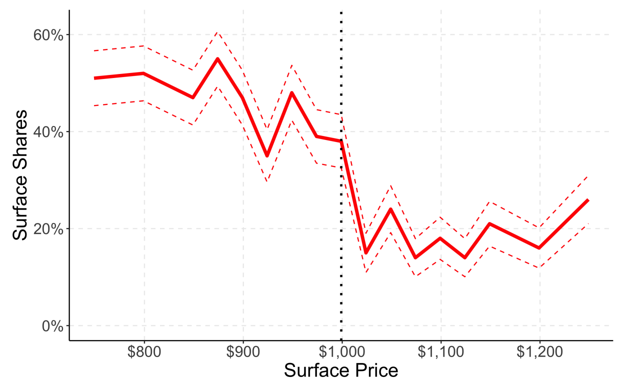
[10] The vertical dotted line in the figure indicates the price of the reference good hereafter.

from this figure. First, the demand curve continues to be broadly downward-sloping. Second, demand for the focal good in this setting appears to decline more sharply than in the previous study. When the focal good is even marginally more expensive than the reference good, demand for it drops to nearly zero and remains small for all higher prices, whereas in Figure 1b we saw demand for the more expensive good remain strictly positive even when the price difference was substantial. This pattern is consistent with perceived horizontal differentiation being higher for laptops than for toothpaste. It is also consistent with prospect theory, which suggests that customers are often more averse to price increases for lower-priced items, holding the percentage price increase constant.

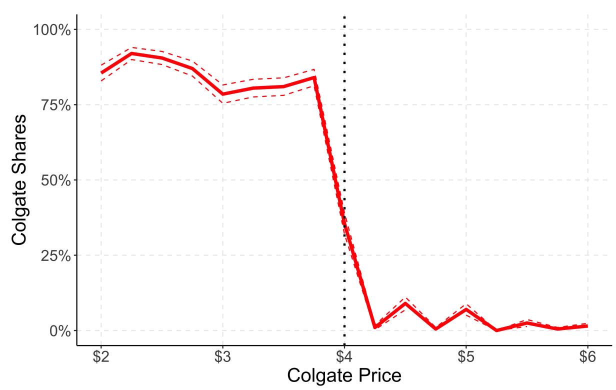
Figure 1: Downward-Sloping Demand Curve



(a) Single Laptop Option



(b) Two Laptop Options

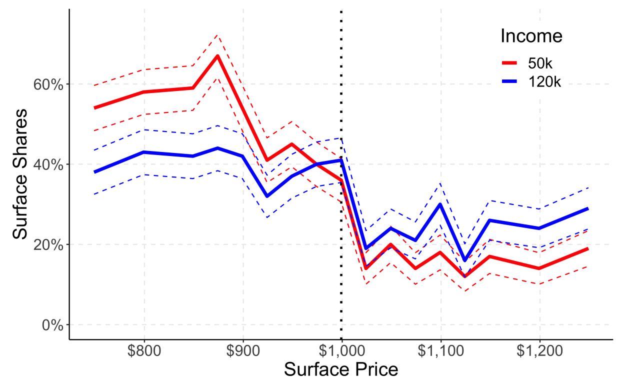


(c) Two Toothpaste Options

3.1.2 Study 2: Impact of income on demand

Economic theory and empirical work suggest that higher-income customers are less sensitive to changes in price than lower-income customers. In order to test whether GPT exhibits this property, we explore the impact of changing the stated level of income of the customer in our prompt. In Section 3.1.1, we prompted GPT with an annual income of $70,000, representative

Figure 2: Impact of Income on Demand



of a median household.[11] In Figure 2, we submit the same set of prices and products to GPT as before (Figure 1b), but change the income to $50,000 or $120,000. We expect a typical demand curve to be flatter for higher-income customers, and this is indeed what we see. When the focal good's price exceeds that of the reference good by $250, lower-income GPT customers reduce their demand for the focal good to roughly 15% while higher-income GPT customers continue to choose the focal good about 25% of the time. In Appendix Section D we show that this result is robust to choosing different sets of laptops and reference prices.

3.1.3 Study 3: State dependence

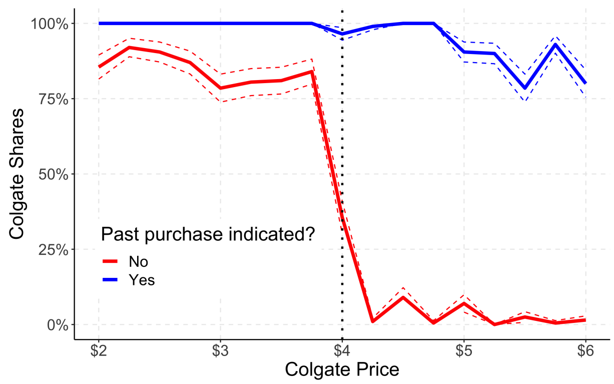
Our next study is designed to test whether GPT's responses are consistent with state dependence. A significant body of work in industrial organization and marketing has studied the magnitude and causes of serial correlation in customers' choices in a variety of contexts. In health insurance markets, for example, inertia drives patients to repeatedly choose the same plan even when the plan's attributes change significantly (e.g., Handel, 2013; Pakes et al., 2021). A number of papers discuss the tools and data necessary to distinguish between various forms of state dependent choice and preference heterogeneity in the choice of consumer packaged goods (Dubé et al., 2010; Levine & Seiler, 2022).

This study offers a choice between two brands of toothpaste, similar to before (Figure 1c), except now we include a phrase in the prompt indicating the brand that the customer purchased previously. In each of our figures we also include the original result with no state-related prompting for reference. In Figure 3a, we add the phrase "The customer says that last time they shopped for toothpaste they purchased the Colgate whitening toothpaste." This modification changes the resulting responses dramatically. When the focal good (Colgate) is cheaper than

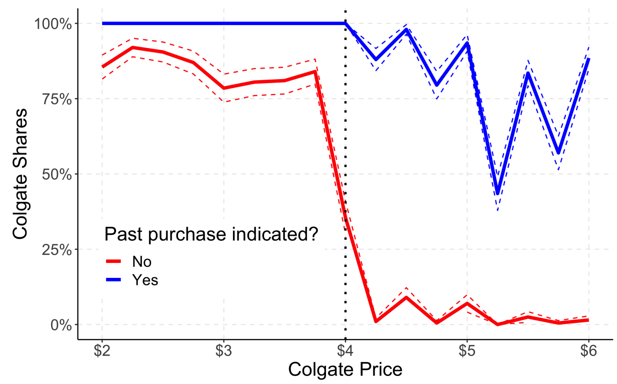
[11] See: https://www.census.gov/library/publications/2022/demo/p60-276.html, accessed March 6, 2022.

the reference good, we find that all responses choose Colgate. Only when Colgate becomes more expensive than the alternative does the demand curve begin to slope downward. Our second exercise highlights the nuances of prompting GPT for our studies. Figure 3b plots the demand curve generated by a prompt that induces state dependence with the alternative phrase "This customer bought the Colgate whitening toothpaste last time they shopped for toothpaste." In this figure, we see a demand curve that, while less monotonic, has a significantly steeper slope than in Figure 3a. Together, these figures indicate that GPT is quite responsive to the structure and content of the response, and that it can use contextual information in intuitive ways to modify the choices it returns.

Figure 3: State Dependence: Previous Colgate Purchase



(a) State prompted as "...customer says"



(b) State prompted as fact

3.1.4 Study 4: Diminishing marginal utility

We also examine whether GPT's responses reflect a diminishing marginal utility of consumption. Within market research, the extent to which marginal utility diminishes is useful for setting quantity discounts, demand forecasting, and inventory management. We modify our prompt with a statement indicating that the randomly selected customer being surveyed has already purchased the good in the past and has x units of the good at home, while varying the value of x. In this exercise, we focus on yogurt, which is often purchased in packs of four to twelve six-ounce cups, and for which customers may both have a stock at home but also consider purchasing more. For each value of x, we ask GPT to provide the customer's willingness to pay for an additional unit of yogurt at the store.

Figure 4a presents our results in the form of a box plot. For each value of x, we present the average (full dots) and median (lines) of the stated willingness to pay. From $x = 0$ to $x = 1$ the mean reported WTP declines sharply. For all x between 1 and 10, the values are about similar. The right-most columns of our plot sets much larger values of x, ranging from 20 to

1,000. Marginal utility diminishes most clearly in these columns. The median customer with 20 units of yogurt at home is willing to pay approximately 15% less than a customer with 10 or fewer pre-purchased units. The average WTP also decreases in the number of units at the 20 to 1,000 units range. While the average WTP at 1,000 units ($3.06) is lower than at 50 units ($3.15) and 100 units ($3.13), the differences are smaller than expected.

Because GPT may infer additional information (e.g., stockpiling, bundling, quantity discounts) from the fact that the customer has x units at home, next we turn to replicate our diminishing marginal utility findings in an *immediate* consumption context: beverage consumption at a restaurant. In this exercise, we focus on glasses of beverages (soda and wine), and ask a "random restaurant goer" how much they would be willing to pay for an additional glass after having ordered and consumed x glasses. Figures 4b and 4c present the results. Unlike the yogurt scenario, we do not find evidence of diminishing marginal returns in these scenarios. We recognize that using consumer surveys to isolate specific relationships (e.g., utility of money or diminishing marginal utility) is challenging, as it is not the usual objective of survey instruments. For example, prompting that a customer has consumed six glasses of wine may not only tell GPT about the customer's prior consumption but also that the customer *really* likes wine. Hence, in addition to the possibility that GPT does not simulate diminishing marginal utility, it is also possible that other factors impact the relationship between consumption levels and willingness to pay for the marginal unit.

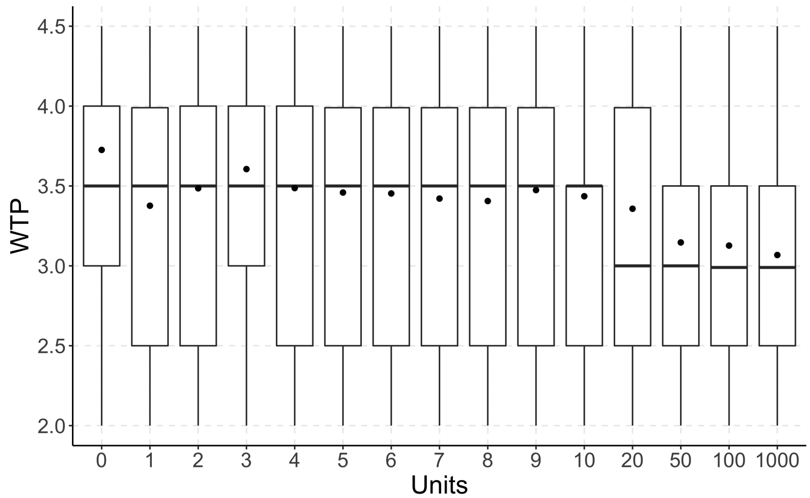
Note that the method by which we have tested diminishing marginal utility is one that is uniquely suited to LLMs. Many customer surveys use conjoint analysis is used to map customers' choices over bundles into estimates of their preferences, which are then used to estimate willingness to pay for products or product attributes. We will return to the approach in Section 3.2.

3.2 Contextualizing GPT Responses

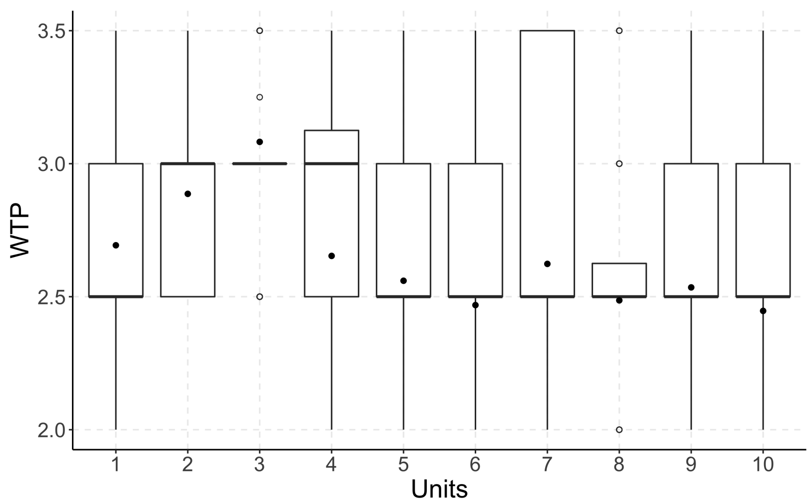
The previous section demonstrates that GPT's answers largely conform to predictions from economic theory and well-documented behavioral patterns. Most of these predictions, however, focus on identifying the correct slope of survey responses without addressing the realism of the responses.

To examine the realism of responses we focus on the willingness-to-pay (WTP) metric. First, we demonstrate that the distribution of WTP for products generates reasonable values for multiple categories of goods. One potential concern with this approach is that rather than a coherent WTP metric, the output reflects the LLM's attempt to match the distribution of prices listed on

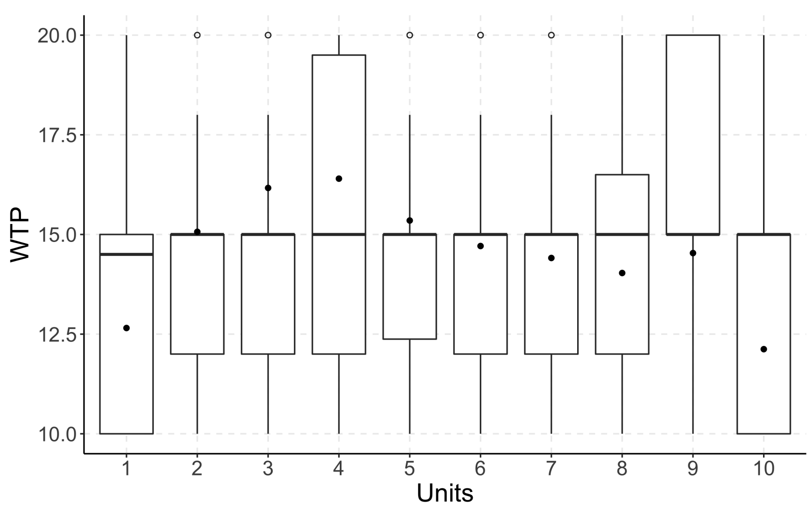
Figure 4: Diminishing Marginal Utility of Consumption



(a) Yogurt



(b) Glass of Soda at a Restaurant



(c) Glass of Wine at a Restaurant

websites selling the goods in the prompt. We address this concern by examining whether we can also back out GPT's willingness-to-pay for product *attributes*, which is much less likely to be stated directly on product pages or in reviews. We use three approaches to solicit GPT's WTP for product attributes: a direct approach (asking for the WTP), an indirect approach (comparing demand functions with and without the attribute), and using a conjoint-style survey of GPT. Importantly, we demonstrate that our results are consistent with those of Fong et al. (2023), who conducted surveys of real individuals and showed that their results matched real market outcomes.

3.2.1 Study 1: Recovering realistic WTP For products

We start by exploring whether asking GPT directly for willingness-to-pay (WTP) for certain products provides a realistic distribution of prices, both for categories which are commonly sold via the internet (laptops, toothpaste) and others which are not (beverages at a restaurant). Figure 5 reports our results. We begin by plotting the distribution of WTP for the Surface Laptop 3 we used in earlier sections (see Figure 5a). The specifications are the same as in earlier sections, but here we do not include any price. The median implied WTP for the Surface Laptop 3 is $1,000, similar to its market price.[12]

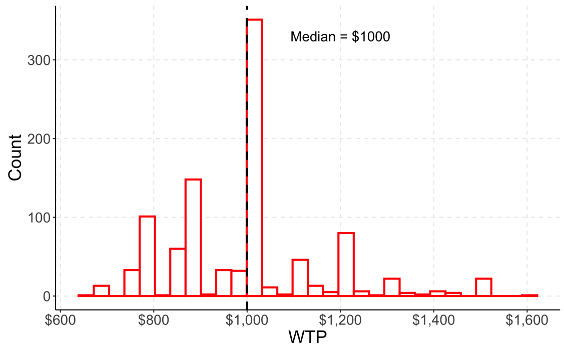
Next, we use more general descriptions to elicit WTP for a good, rather than WTP for a particular brand. Recall from section 3.1.4 where we collected the willingness to pay for yogurt, that the range of prices and the mean and median were appropriate for the product category. In Figure 5b we use "whitening toothpaste" to solicit the distribution of prices. We then move on from packaged goods that can be purchased online, and examine willingness to pay in another context — a restaurant. We ask for the WTP for a glass of soda (Figure 5c) and a glass of wine (Figure 5e) at a restaurant, and find that the median WTP for wine is six times higher than for soda (median of $15 compared to $2.5). Finally, we also demonstrate that the WTP for a soda can at a supermarket is lower than at a restaurant with a median of $2 (see Figure 5d) and that the WTP for a bottle of wine at a restaurant is more than 3 times the WTP for a glass of wine (Figure 5f).

3.2.2 Study 2: Recovering realistic WTP For attributes

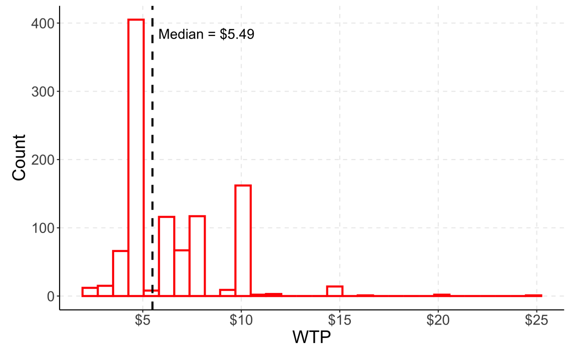
After demonstrating that asking GPT directly for WTP provides a realistic distribution of values for a variety of goods, we turn to examining whether we can recover estimates of WTP for attributes from GPT's responses. We demonstrate our results using two examples: fluoride (at-

[12]A sample of responses that GPT provided to this prompt appears in Appendix Section C.

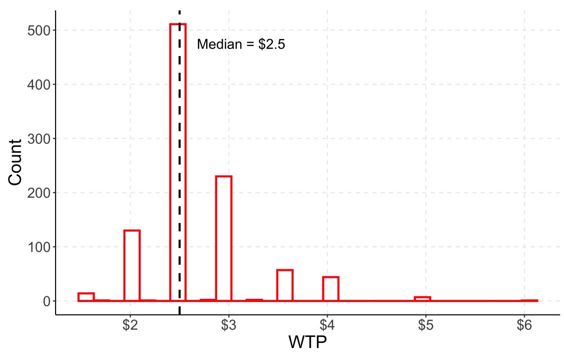
Figure 5: Willingness to Pay for Products



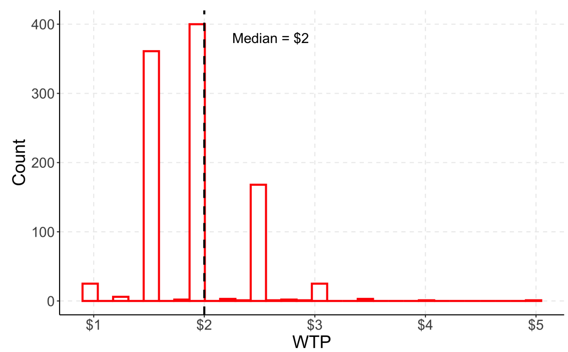
(a) Surface Laptop 3



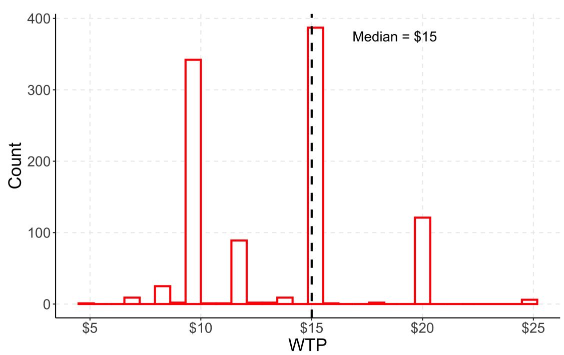
(b) Whitening Toothpaste



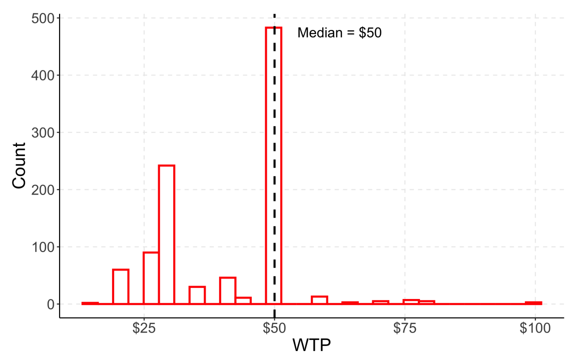
(c) “Glass of soda (at restaurant)"



(d) “Can of soda (at supermarket)"



(e) “Glass of wine (at restaurant)"

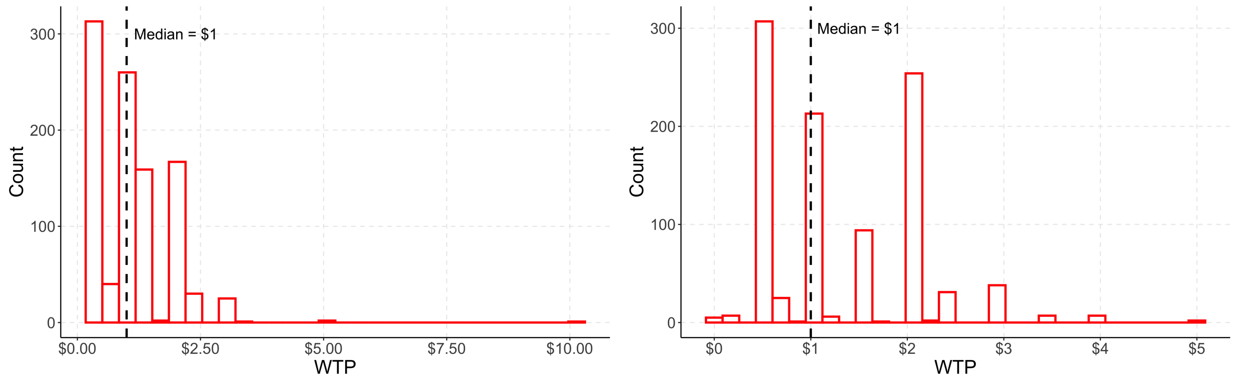


(f) “Bottle of wine (at restaurant)"

tribute) in toothpaste (product), and aluminum (attribute) in deodorant (product), examples we borrow from Fong et al. (2023).

We utilize three different strategies: a direct solicitation approach, an indirect solicitation approach, and a conjoint approach. We note that a direct solicitation of WTP for product attributes is not the typical method for traditional human-based market research, as humans' ability to quantify these measures is limited. Marketing researchers and practitioners often use conjoint analysis to derive WTP from survey responses. However, GPT may be more so-

Figure 6: Willingness to Pay for Attributes – Direct Solicitation



(a) Fluoride in toothpaste

(b) No aluminum in deodorant

phisticated than humans and may be able to calculate or infer WTP for attributes from other training data. Therefore, we test the ability of GPT to provide WTP for attributes directly or in a relatively simple indirect method before moving on to a conjoint study.

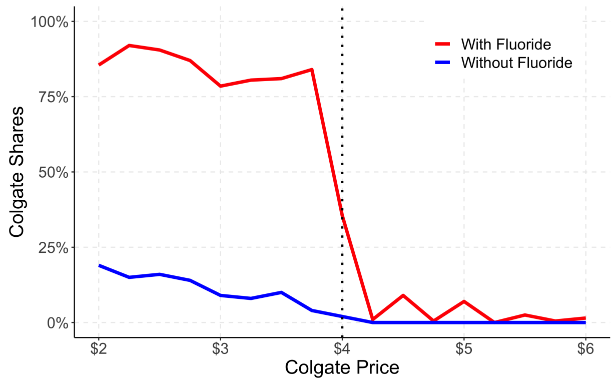
Direct solicitation For the direct solicitation approach, we offer two identical goods that differ only on the existence of the attribute of interest. For toothpaste, we offer Colgate whitening toothpaste and ask GPT how much more it would be willing to pay for the option with fluoride over the option without fluoride. Similarly, we ask GPT how much more it would be willing to pay for a Dove scented deodorant without aluminum compared to a version that contains aluminum.[13] Figure 6a provides the distribution of WTP for fluoride using this approach, and Figure 6b provides the WTP for "no aluminum." The median and average WTP is $1.00 and $1.20 for fluoride and $1.00 and $1.30 for "no aluminum."

Indirect solicitation The indirect solicitation approach consists of two steps. First, we use GPT to estimate the demand for the good with and without the attribute using a similar paradigm that we used in Section 3.1.1 to generate two demand curves. Then, we compare the demand curves to derive WTP for the attribute.

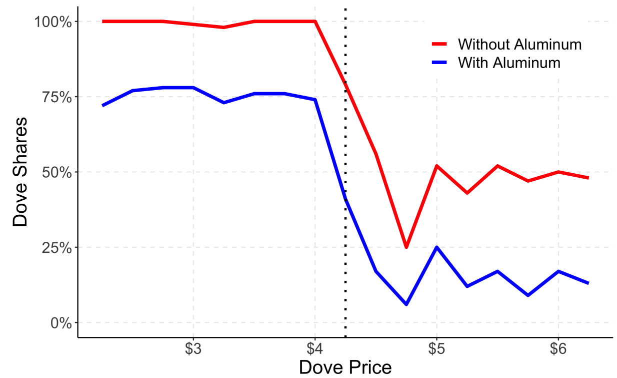
For toothpaste, we first estimate the demand for Colgate whitening toothpaste without fluoride to generate a demand curve (this time, the focal good did not have fluoride, but the reference good priced at $4 did have fluoride). Then, we compare the demand curve for the toothpaste with fluoride that we showed earlier (from Figure 1c) with the demand curve for the toothpaste without fluoride (see Figure 7a for the resulting demand curves) to derive WTP for fluoride. At each price p on the "without fluoride" demand curve, we calculate the price p' such that

[13] In WTP queries, we query GPT 1,000 times for each product.

Figure 7: Willingness to Pay for Attributes – Indirect Solicitation



(a) Toothpaste, Demand curve



(b) Deodorant, Demand curve

demand for toothpaste with fluoride at p' is equal to the demand for flouride-less toothpaste at p. Our WTP measure is then $p' - p$, which amounts to taking horizontal differences between the demand curves in Figure 7a. For example, when the price of Colgate without fluoride is $p = \$3.00$, the shares are 9%, which corresponds to the Colgate with fluoride shares at a price of $p' = \$4.19$, implying a WTP for fluoride of \$1.19. We note that this is comparable to the median WTP of \$1.00 generated by the direct elicitation approach in Figure 6a.

We follow the same approach for deodorant (using the Speed Stick brand with a price of \$4.25 as a reference price, based on Fong et al. (2023)), which generates the curves in Figure 7b. As can be seen in the figure, aluminum-free deodorant is preferred by GPT over deodorant containing aluminum. The implied WTP for aluminum is -\$1.26 when deodorant is priced at \$3.00.

Also implied by these figures are GPT's brand preferences. In Figure 7a, it seems that GPT has a preference of Crest over Colgate: when both brands have fluoride, the shares for Colgate are below 100% even when Colgate is significantly cheaper than Crest, and at price parity the shares of Colgate are lower than those of Crest. Similarly, Figure 7b reflects that GPT prefers Dove over Speed Stick. These preferences are consistent with those at Fong et al. (2023), who report a slight preference for Crest (the observed market shares they report are 34.7% for Crest versus 33% for Colgate in the toothpaste category) and a significant preference for Dove (the reported shares are 46% for Dove versus 22% for Speed Stick in the deodorant category).

Recovering preferences via conjoint For our final approach, we recover preferences using the conjoint analysis paradigm. Conjoint is widely used in industry and academia for estimating customer WTP, and has been shown to be able to uncover customer preferences for different product attributes jointly (see Green and Rao (1971), Green and Srinivasan (1978), and Green and Srinivasan (1990) for a review).

We evaluate GPT on two dimensions. First, we treat the responses from GPT as if they were from randomly chosen consumers and test whether the effects of changes in price and non-price attributes on choice probabilities are consistent with economic predictions. Although we have shown above that GPT's demand is decreasing with a product's own price when the other price is fixed, our studies in this section test whether the same holds true when prices and attributes of both goods can vary across prompts. Second, we use the queried responses to estimate a multinomial logit model, in order to evaluate the realism of model-based estimates of WTP.

The use of toothpaste and deodorant choices is inspired by Fong et al. (2023), who run a conjoint study and confirm that their experimental estimates are consistent with market outcomes. We use the brands we used in our earlier analyses, Colgate and Crest for toothpaste and Dove and Speed Stick for deodorant. Similar to Fong et al. (2023), we use three levels of prices for each of the goods[14], and two attribute levels for fluoride and aluminum (with, without).

Typical conjoint studies involve generating choice sets that are orthogonal across configurations and balance attributes across choices. Study participants are then presented with 10–15 scenarios comparing 2–3 products (as well as a no-purchase option) which are a subset of the full set of configurations. Because we are not limited by humans' time or their ability to process complex information, we choose to create the full set of options for each brand: three price levels for each of the attribute options, yielding a total of 36 configurations. We collect 300 responses for each configuration, for a total of 10,800 responses. Using GPT, each conjoint study with 10,800 responses cost $3 and took roughly 35 minutes to run. These are negligible costs, particularly compared to typical fees charged for commercially-provided conjoint studies.

The overall results seem somewhat consistent with the aforementioned market shares in terms of the strength of brand preferences. For toothpaste, 2,300 responses chose Crest, 2,468 chose Colgate, and the remainder chose the no-purchase option. For deodorant, 1,207 chose Dove, 501 chose Speed Stick, and the remainder chose the no-purchase option.

We first present simple regressions and separate estimates for each brand's price and fluoride attribute using ordinary least squares (Table 1). Reassuringly, the estimates have the expected signs: when Crest (Colgate) is priced higher, the likelihood of choosing Colgate (Crest) is

[14] For toothpaste we use: $0.99, $1.99, $2.99, and for deodorant we use: $1.99, $2.99, $3.99.

higher; when Crest (Colgate) includes fluoride, the likelihood of choosing Crest (Colgate) is higher, and the likelihood of choosing Colgate (Crest) is lower.

Table 1: Toothpaste Conjoint: Choice Determinants

	$\mathbb{I}$(*Choice=Colgate*)	$\mathbb{I}$(*Choice=Crest*)
	(1)	(2)
Crest price	0.120*** (0.004)	−0.081*** (0.005)
Colgate price	−0.080*** (0.004)	0.134*** (0.005)
$\mathbb{I}$(Crest fluoride)	−0.199*** (0.007)	0.307*** (0.008)
$\mathbb{I}$(Colgate fluoride)	0.251*** (0.007)	−0.210*** (0.008)
Constant	0.122*** (0.014)	0.142*** (0.015)
Observations	10,800	10,800

Significance level: 10% (*); 5% (**); 1% (***).

In the deodorant case (Table 2), the aluminum content of Speed Stick has no impact (the coefficient is negative but indistinguishable from zero) on the choice of Dove. The other coefficients are in the expected direction.

Table 2: Deodorant Conjoint: Choice Determinants

	$\mathbb{I}$(*Choice=Dove*)	$\mathbb{I}$(*Choice=Speed Stick*)
	(1)	(2)
Dove price	−0.060*** (0.004)	0.033*** (0.002)
Speed Stick price	0.040*** (0.004)	−0.032*** (0.002)
$\mathbb{I}$(Dove Aluminum)	−0.047*** (0.006)	0.022*** (0.004)
$\mathbb{I}$(Speed Stick Aluminum)	−0.001 (0.006)	−0.046*** (0.004)
Constant	0.197*** (0.016)	0.055*** (0.011)
Observations	10,800	10,800

Significance level: 10% (*); 5% (**); 1% (***).

Our second set of results are estimates from a multinomial logit choice model, estimated by treating GPT's responses as if it were generated by a random sample of customers. We report the results for toothpaste in Columns 1 and 2, and for deodorant in Columns 3 and 4. Based on these estimates, the implied WTP for fluoride in our sample is \$3.40 (calculated by dividing the fluoride coefficient by the absolute value of the price coefficient). Our estimates from this model are substantially larger than in the preceding sections and are quite similar to the estimates in Fong et al. (2023), who conduct a real-world conjoint to estimate customer preferences for toothpaste and estimate the WTP for fluoride to be \$3.27. Table 3 includes both the multinomial logit results (Column 1) and a random coefficient model (Column 2). The results are consistent across both estimation methods (in the random coefficient model our WTP estimate is $3.30). The implied WTP for aluminum in our sample is -$0.99 (-$0.92 in the random coefficient model), compared to the estimates of -$1.97 (-$1.53) in Fong et al. (2023), which are closer to the estimates we obtained in the preceding sections.

Table 3: Conjoint Results

	Toothpaste		*Deodorant*	
[-1.8ex]	(1)	(2)	(3)	(4)
Price	-0.484^{***} (0.021)	-0.504^{***} (0.034)	-0.692^{***} (0.034)	-0.762^{***} (0.057)
Attribute	1.647^{***} (0.037)	1.662^{***} (0.044)	-0.685^{***} (0.054)	-0.697^{***} (0.058)
Brand 1 Dummy	-0.801^{***} (0.051)	-0.778^{***} (0.060)	0.229^{**} (0.099)	0.354^{***} (0.129)
Brand 2 Dummy	-0.491^{***} (0.050)	-0.457^{***} (0.063)	-0.678^{***} (0.103)	-0.556^{***} (0.125)
σ Price		0.155^{**} (0.067)		0.156^{***} (0.059)
σ attribute		1.049^{***} (0.149)		0.075 (0.271)
Observations	10,800	10,800	10,800	10,800

Significance level: 10% (*); 5% (**); 1% (***).

In Columns (1) and (2) Attribute="fluoride", and in Columns (3) and (4) Attribute="aluminum"; Brand dummies are "Colgate" (1) and "Crest" (2) for toothpaste, and "Dove" (1) and "Speed Stick" (2) for deodorant.

Overall, these conjoint-style exercises demonstrate that querying GPT can generate estimates that are similar to those generated by human-based conjoint studies. In other words, we are able to recover realistic and consistent aggregate "utilities" from GPT responses to choice prompts. These results are highly encouraging, especially considering that GPT is trained to merely generate text that is likely to be encountered and not to rationalize choices based on attribute characteristics.

4 Guidelines and Limitations in Querying GPT

Designing and running the studies in this paper has allowed us to identify some simple guidelines that improve the quality of the responses given by GPT, as well as important cases in which GPT exhibits particular sensitivity or unreliability. Like most applications of GPT, we have found prompt engineering to be important for retrieving a useful response from GPT. We offer the examples we came across below, while recognizing that these are a small representation of a full set of guidelines for using GPT in market research.

Sensitivity to Response Order. When offered multiple options, GPT is significantly more likely to choose the option that is listed first. For all of our results that include two options, we randomize the order of these options, and run the surveys with one option appearing first for half of our sample.

Inducing Choosing the Outside Option. The fraction of GPT survey responses in which the

GPT customer chooses one of the available options (rather than choosing not to purchase) depends on the precise phrasing of the prompt. Consider the following two potential phrases to include in the prompt after describing the available choices:

- "They also have the option not to purchase a laptop. The customer is asked, after they finish shopping: Which laptop, if any, did you purchase?"
- "They also have the option not to purchase a laptop. The customer is asked, after they finish shopping: Did you purchase a laptop? If so, which one?"

Although their meaning is quite similar, in practice we find that the first phrase yields only a handful of responses in which the outside option is chosen, while the second phrase leads to outside option shares of roughly 30% to 60%. Importantly, conditional on making a choice, the implied marked shares are similar between the two types of phrases. We see a similar pattern arise when, earlier in our prompt, we specify that the customer "sees two options," rather than stating that the customer "has three options," which explicitly includes the outside option and results in more realistic market shares. These differences in prompting were especially crucial for two of our exercises: the binary choice study, and the conjoint study. For these studies, we used the language "Did you purchase... If so, which one?"

Specificity in requested output. We found GPT to be verbose in its responses to our early prompts. For example, if we ask a question aimed at eliciting willingness to pay, (e.g., "What is the maximum price you would be willing to pay for X?") we were likely to receive an essay-like response, which includes the reasoning for the answer, or a range of prices. Alternatively, requesting a single price as an answer was more likely to produce a single price and a more concise response overall. GPT responses are sensitive as well to the exact framing of such a prompt. For example, when the prompt included "Please answer by giving an amount in dollars" GPT only provided round dollar amounts, whereas specifying "amount in dollars and cents" led to the expected output.

An interesting question remains as to which of these guidelines and limitations are inherent to querying LLMs, and which are artifacts of surveying consumers that are merely carried over by GPT — is GPT sensitive to response order because it is an LLM or because humans tend to select the first option more frequently (Ferber, 1952)? This is just one of many exciting questions that we anticipate future research in this area will address.

5 Conclusion

Our results suggest that GPT, and LLMs more broadly, can serve as a powerful tool for uncovering customer preferences. Our first set of studies highlights that when prompted as if it were a randomly selected customer, GPT exhibits a number of behaviors that are consistent with economic theory, including both declining price sensitivity with income and state dependence. These two properties are particularly notable because of their complexity. Not only are the resulting demand curves elicited from GPT downward-sloping, but they are coherent across the different queries and scenarios (i.e, incomes, historical purchase behavior), and the relationship between different demand curves is similarly coherent. This is an essential feature for these types of systems to be used as tools for marketing researchers and practitioners.

Our second set of results demonstrates that the estimates derived from GPT's responses are realistic and consistent with values obtained from existing research. We begin by directly soliciting WTP for multiple categories of goods and find that responses are of realistic magnitudes, while noting that such direct solicitation of WTP from human subjects is known to suffer from many shortcomings (Wertenbroch & Skiera, 2002). We then demonstrate that GPT can generate richer results that marketing researchers and practitioners may find interesting. When given a conjoint-style survey, GPT is able to generate estimates of WTP for fluoride and deodorant that are close to existing research, and exhibits substitution patterns that are often expected from real consumer choice data, including correct signs of own- and cross-price effects and substitution on non-price attributes. To us, these results suggest that GPT could also be useful also for the development of new products. Standard empirical approaches for predicting demand for new products rely on estimating customer preferences on characteristic space and extrapolating to new combinations of attributes. While we do not expect GPT to be able to predict the future any better than humans can, we do believe that it can return preferences in ways that may allow researchers to forecast demand for new goods using existing methods.

Just as there is a substantial literature discussing the best ways to solicit customer preferences for goods, we envision a similar literature will evolve for GPT-based surveys. Our results suggest significant potential for progress in translating GPT's capabilities into practical tools for researchers and businesses. At present, we see three paths forward. First, with minimal effort beyond the types of prompting discussed in this paper, GPT can serve as a realistic simulator of customer choice. Before running a conjoint, or prior to running code on a new data set, a researcher can prompt GPT to generate artificial data. Given the emergent properties highlighted here, the results may be more realistic than standard approaches for generating simulated choice data.

Second, users may adapt GPT to suit specific contexts by providing it with various forms of "knowledge." In our simple exercises, we give GPT details about income and prior purchases. One might imagine assigning GPT "personas," preferences, and product or market information of increasing complexity. Moreover, researchers may wish to build a distribution over such inputs according to the empirical distribution that reflects the target customers. For example, to generate a nationally representative sample, one could inject incomes into the prompt, drawing from the full national income distribution. In a different context, a researcher may have convincing prior knowledge about some moment of the WTP distribution and can incorporate it to conduct calibration exercises, where the prompt design is engineered to match the known moment.

Finally, we expect that LLMs will become more useful for market research in the future, parallel to the rapid improvement in the sophistication of these models. As LLMs improve in accuracy (as widely reported from the release of GPT-4) and access more data (as demonstrated by their use in popular search engines), we are optimistic that their ability to absorb and infer rich aspects of consumer behavior will likewise increase. While we appeal to established market research paradigms to illustrate the usefulness of GPT as a source of truth, LLMs may give rise to new market research paradigms unbounded by the limits of human subjects research.

Importantly, we also offer some words of caution. Much work needs to be done to evaluate which market research objectives LLMs are best suited to, and for which ones they are a poor substitute for existing methods. We have identified a few areas in which GPT appears to fall short of capturing preferences, such as its minimal ability to reflect diminishing marginal utility. We expect that there are at least a few more. For instance, because GPT is "pre-trained," without additional training data provided by the researcher or access to the internet, it may reveal static preferences. Additionally, our work emphasizes the sensitivity of GPT to how prompts are worded (see Section 4 which provides examples we came across, although our results were overall robust to different wordings).

Managers and researchers should also be aware that LLMs are known to occasionally "hallucinate" and return incorrect information; would they similarly hallucinate a prediction of success for a new product or feature? Such questions are critical for establishing the usefulness of LLMs for key market research objectives. Our results that GPT provides different responses based on our inputs (such as income, prices, brands, product attributes) are encouraging, and suggest that GPT does not provide completely random (or "garbage") responses. Instead, it adapts to the input in a way that is consistent with theoretical predictions and with human behavior. We therefore cautiously believe it could be used for market research, with a critical eye examining the outputs and their correctness.

Finally, while we see GPT as a means for managers and researchers to uncover preferences in lieu of survey-based or observational methodologies, we believe that disclosure of the source of inferences from GPT is necessary both from an ethical and an external validity standpoint.

References

Aher, G., Arriaga, R. I., & Kalai, A. T. (2022). Using large language models to simulate multiple humans. *arXiv preprint arXiv:2208.10264*.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2022). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 1–15.

Burnap, A., Timoshenko, A., & Hauser, J. R. (forthcoming). Product aesthetic design: A machine learning augmentation. *Marketing Science*.

Dubé, J.-P., Hitsch, G. J., & Rossi, P. E. (2010). State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics*, *41*(3), 417–445.

Ferber, R. (1952). Order bias in a mail survey. *Journal of Marketing*, *17*(2), 171–178.

Fong, J., Guo, T., & Rao, A. (2023). Debunking misinformation about consumer products: Effects on beliefs and purchase behavior. *Journal of Marketing Research*.

Gerstner, E. (1985). Do higher prices signal higher quality? *Journal of marketing research*, *22*(2), 209–215.

Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research*, *8*(3), 355–363.

Green, P. E., & Srinivasan, V. (1978). Conjoint Analysis in Consumer Research: Issues and Outlook. *Journal of Consumer Research*, *5*(2), 103–123.

Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, *54*(4), 3–19.

Handel, B. R. (2013). Adverse selection and inertia in health insurance markets: When nudging hurts. *American Economic Review*, *103*(7), 2643–2682.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*.

Johnston, R. J., Boyle, K. J., Adamowicz, W., Bennett, J., Brouwer, R., Cameron, T. A., Hanemann, W. M., Hanley, N., Ryan, M., Scarpa, R., et al. (2017). Contemporary guidance for stated preference studies. *Journal of the Association of Environmental and Resource Economists*, *4*(2), 319–405.

Kroes, E. P., & Sheldon, R. J. (1988). Stated preference methods: An introduction. *Journal of Transport Economics and Policy*, 11–25.

Levine, J., & Seiler, S. (2022). Identifying state dependence in brand choice: Evidence from hurricanes. *Marketing Science*.

Li, P., Castelo, N., Katona, Z., & Sarvary, M. (2023). Language models for automated market research: A new way to generate perceptual maps. *Available at SSRN 4241291*.

Mollick, E. R., & Mollick, L. (2023). Using ai to implement effective teaching strategies in classrooms: Five strategies, including prompts. *Available at SSRN 4391243*.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, *31*(3), 521–543.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pakes, A., Porter, J. R., Shepard, M., & Calder-Wang, S. (2021). *Unobserved heterogeneity, state dependence, and health plan choices* (tech. rep.). National Bureau of Economic Research.

Payne, J. W. (1982). Contingent decision behavior. *Psychological Bulletin*, *92*(2), 382–402.

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv preprint arXiv:2302.06590*.

Schindler, R. M., & Kirby, P. N. (1997). Patterns of rightmost digits used in advertised prices: Implications for nine-ending effects. *Journal of Consumer Research*, *24*(2), 192–201.

Thomas, M., & Morwitz, V. (2005). Penny wise and pound foolish: The left-digit effect in price cognition. *Journal of Consumer Research*, *32*(1), 54–64.

Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, *38*(1), 1–20.

Wertenbroch, K., & Skiera, B. (2002). Measuring consumers' willingness to pay at the point of purchase. *Journal of Marketing Research*, *39*(2), 228–241.

Appendix

A Code Example

The following Julia code was used to collect GPT responses for Figure 1a:

```
# Import packages (intall first if needed)
using CSV, JSON, OpenAI, DataFrames

sleep_time = 5;
length_per_iter = 50;
# Include api_key provided by OpenAI service
api_key = "";
# Vectors of incomes and prices
incomes = ["\$50k", "\$70k", "\$120k"];
prices_int = [749,799,849,874,899,924,949,974,999,1024,1049,1074,1099,1124,1149,1199,1249];
prices_string = "\$".*string.(prices_int);

# -------------------------------------------------------------
# All functions for prompting, extracting responses, etc.
# -------------------------------------------------------------
function query(prompt, N)
    rvec = [];
    if N <=128
        rvec = create_completion(api_key, "text-davinci-003";
            prompt= prompt,
            temperature=1,
            max_tokens=100,
            top_p=1.0,
            n = N,
            frequency_penalty=0.0,
            presence_penalty=0.0,
            stop=["\"\"\""])
    else
        nqueries = floor(N/128)+1;
        if floor(N/128) == N/128
            nqueries = floor(N/128);
        end
        for i = 1:nqueries
            if N > (i-1)*128
                println("Waiting before/between queries.......")
                sleep(sleep_time)
                n_for_request = 128;
                if (i==nqueries) & (N > (i-1)*128)
                    n_for_request = Int(N - (i-1)*128);
                end
                rtemp = create_completion(api_key, "text-davinci-003";
                    prompt= prompt,
                    temperature=1,
                    max_tokens=30,
                    top_p=1.0,
                    n=n_for_request,
                    frequency_penalty=0.0,
                    presence_penalty=0.0,
                    stop=["\"\"\""])
                println("Query $(i)/$(nqueries) Done")
                push!(rvec, rtemp)
            end
        end
    end
    return rvec
end

function get_choices(response_vec)
    choices = [];
    if typeof(response_vec) <: OpenAIResponse
        nresponses = length(response_vec.response.choices);
        for i = 1:nresponses
            push!(choices, getindex(response_vec.response.choices,i)[:text]);
```

```
        end
    else
        for outer_i = 1:length(response_vec)
            nresponses = length(response_vec[outer_i].response.choices);
            for i = 1:nresponses
                push!(choices, getindex(response_vec[outer_i].response.choices,i)[:text]);
            end
        end
    end
    return choices
end

function make_prompt_surface(; income = "\$70k", surface_price = "\$999")
    prompt = """A customer is randomly selected while shopping for laptops. Their annual income is $(income).
    While shopping, the customer sees a Surface Laptop 3, Price: $(surface_price), Processor: Intel Core i5, RAM:
         8GB, Screen Size: 13.5in, SD: 128GB
    The customer is asked, after they finish shopping: Did you purchase any laptop? If so, which one?
    Customer: """
    return prompt
end

# ----------------------------------------------------------------
# Study 1: Downward sloping demand curve - laptop
# ----------------------------------------------------------------
response_vec = [];
price_vec = [];
choices_vec = [];
N = 300;
foo = collect(Iterators.product(prices_string, incomes));
foo = repeat(foo, trunc(Int, N/length_per_iter));
foo = reshape(foo, :, 1);

while length(foo) != 0
    surface_prompt = make_prompt_surface(income=foo[1][2], surface_price=foo[1][1]);
    responses = query(surface_prompt, length_per_iter);
    choices = get_choices(responses);
    df = DataFrame(income=foo[1][2], price=foo[1][1], choice=choices);
    CSV.write("study_1a.csv", df, append=true);
    println("price $(foo[1][1]), income $(foo[1][2]) saved. $(length(foo)) left.");
    foo = foo[Not(1), :];
    sleep(sleep_time)
end
```

B Prompts and study details

Below we provide the complete sets of prompts for our analyses. As mentioned in Section 4, whenever we presented two options in a prompt, we ensured to randomize the order of the option. In the interest of clarity and space, we only detail one of those options below.

B.1 Prompts for Section 3.1.1

The following prompts were used to create the data for Figure 1:

- For the single laptop option:

  “A customer is randomly selected while shopping for laptops. Their annual income is $\$income$.

  While shopping, the customer sees a Surface Laptop 3, Price: $\$surfacePrice$, Processor: Intel Core i5, RAM: 8GB, Screen Size: 13.5in, SD: 128GB

  The customer is asked, after they finish shopping: Did you purchase any laptop? If so, which one?

  Customer: "

- For the two laptops:

  “A customer is randomly selected while shopping for laptops. Their annual income is $\$income$.

  While shopping, the customer has three options:

  – Surface Laptop 3, Price: $\$surfacePrice$, Processor: Intel Core i5, RAM: 8GB, Screen Size: 13.5in, SD: 128GB

  – Macbook Air (2019), Price: \$999, Processor: Intel Core i5, RAM: 8GB, Screen Size: 13.3in, SD: 128GB

  They also have the option not to purchase a laptop. The customer is asked, after they finish shopping: Which laptop, if any, did you purchase?

  Customer: ”

- For the two toothpastes:

"A customer is randomly selected while shopping in the supermarket. Their annual income is $\$income$.

While shopping, the customer passes by the toothpaste aisle and sees two options:

– Colgate whitening toothpaste with fluoride, price $\$colgatePrice$.

– Crest whitening toothpaste with fluoride, price \$4.

They also have the option not to purchase toothpaste. The customer is asked, after they finish shopping: Which toothpaste, if any, did you purchase?

Customer: "

B.2 Prompts for Section 3.1.2

For this section, we used the prompt for two laptops from the previous section, while varying the income level.

B.3 Prompts for Section 3.1.3

- State prompted as "customer says":

  "A customer is randomly selected while shopping in the supermarket. Their annual income is $\$income$.

  While shopping, the customer passes by the toothpaste aisle and sees two options:

  – Colgate whitening toothpaste with fluoride, price $\$colgatePrice$.

  – Crest whitening toothpaste with fluoride, price \$4.

  They also have the option not to purchase toothpaste. The customer says that last time they shopped for toothpaste they purchased the Colgate whitening toothpaste.

  The customer is asked, after they finish shopping: which toothpaste, if any, did you purchase this time?

  Customer: "

- State prompted as fact:

  "A customer is randomly selected while shopping in the supermarket. Their annual income is $\$income$.

  While shopping, the customer passes by the toothpaste aisle and sees two options:

  – Colgate whitening toothpaste with fluoride, price $\$colgatePrice$.

  – Crest whitening toothpaste with fluoride, price \$4.

  They also have the option not to purchase toothpaste. This customer bought the Colgate whitening toothpaste last time they shopped for toothpaste.

  The customer is asked, after they finish shopping: which toothpaste, if any, did you purchase this time?

  Customer: "

B.4 Prompts for Section 3.1.4

- For yogurt at the supermarket:

  "A customer is randomly selected while shopping in the supermarket. Their annual income is $\$income$.

  The customer has $\#units$ units of yogurt at home.

  The customer is asked: What is the maximum price you would be willing to pay for one additional unit of yogurt? please give a single price as your answer.

  Customer: \$"

- For beverages at a restaurant, we used the same prompt, replacing $beverage$ with soda and wine:

  "A customer is randomly selected while sitting at a restaurant. Their annual income is $\$income$.

  The customer has ordered and already consumed $number$ of glasses of $beverage$.

The customer is asked: What is the maximum price you would be willing to pay for one additional glass of $beverage$? please give a single price as your answer.

Customer: $”

B.5 Prompts for Section 3.2.1

- Laptop:

  “A customer is randomly selected while shopping for laptops. Their annual income is $$income$.

  While shopping, the customer sees a Surface Laptop 3, Processor: Intel Core i5, RAM: 8GB, Screen Size: 13.5in, Screen Size: 13.5in, SD: 128GB

  The customer is asked: What is the maximum price you would be willing to pay for this Surface laptop? please give a single price as your answer.

  Customer: $"

- Other goods:

  We change the customer location (sitting at a restaurant / shopping in the supermarket), as well as the good, but use the general prompt:

  “A customer is randomly selected while sitting at a restaurant. Their annual income is $$income$.

  The customer is asked: What is the maximum price you would be willing to pay for one glass of wine? please give a single price as your answer.

  Customer: $”

B.6 Prompts for Section 3.2.2

- For direct solicitation (fluoride and toothpaste case):

  “A customer is part of a survey meant to elicit their willingness to pay for different attributes of goods. Their annual income is $$income$.

  The customer is asked to consider two options:

– Option 1: Colgate toothpaste, without fluoride, whitening

– Option 2: Colgate toothpaste, with fluoride, whitening

The customer is then asked: ‘how much more would you be willing to pay for Option 2 than for Option 1?’ Please answer by giving an amount in dollars and cents.

Customer: $”

- For implied demand curve calculation (fluoride and toothpaste case):

  “A customer is randomly selected while shopping in the supermarket. Their annual income is $\$income$.

  While shopping, the customer passes by the toothpaste aisle and sees two options:

  – Colgate whitening toothpaste without fluoride, price $\$colgatePrice$.

  – Crest whitening toothpaste with fluoride, price $4.

  They also have the option not to purchase toothpaste. The customer is asked, after they finish shopping: Which toothpaste, if any, did you purchase?

  Customer: ”

- For the conjoint studies (fluoride and toothpaste case):

  “A customer is randomly selected while shopping in the supermarket. Their annual income is $\$income$.

  While shopping, the customer passes by the toothpaste aisle and sees two options:

  – Colgate whitening toothpaste $colgateFluoride$ fluoride, price $\$colgatePrice$.

  – Crest whitening toothpaste $crestFluoride$ fluoride, price $\$crestPrice$.

  They also have the option not to purchase toothpaste. The customer is asked, after they finish shopping: Did you purchase any toothpaste? If so, which one?

  Customer: ”

B.7 Number of observations collected

For studies in which we provided prices, we collected 300 responses for each price level. For studies in which we explicitly ask for willingness to pay, we ask for 1,000 responses and plot the distribution of those responses.

C Sample responses

C.1 Responses for single laptop option (Figure 1a)

- "Yes, I purchased a Surface Laptop 3."
- "I did purchase the Surface Laptop 3. It had a great processor, good RAM, and the screen size was perfect for my needs. The 128GB SD was a bonus too."
- "No, I did not purchase any laptop. I was looking at the Surface Laptop 3 but ultimately decided it was not the right one for me."
- "No, I did not purchase a laptop. I found the Surface Laptop 3 to be a bit too expensive for my budget."
- "No, I didn't purchase any laptop. I was looking at the Surface Laptop 3, but I wasn't convinced it was a good value considering its specs."

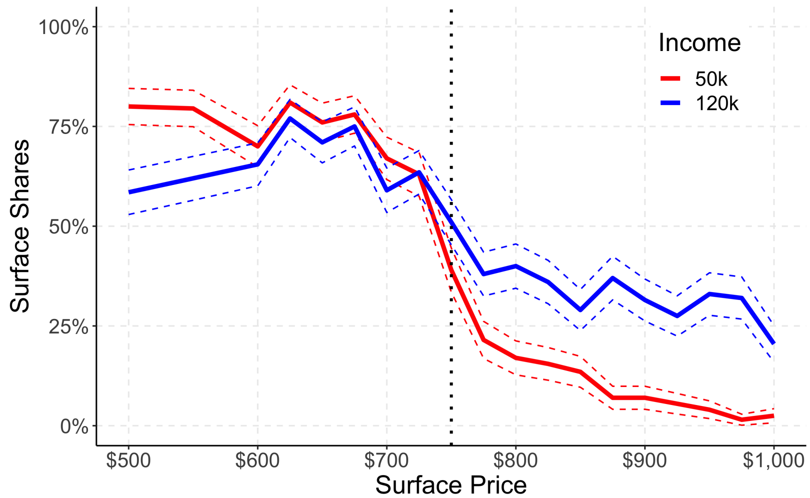
C.2 Responses for WTP question (Figure 5a: Surface laptop)

- "1,249.99"
- "1,400 is the highest I would be willing to pay for this particular laptop."
- "1,500. I think a laptop at this price range would have enough features to meet my needs without breaking the bank."
- "1,500 maximum; I'd be willing to go a bit higher if I can get a discount. I'm looking to get the best laptop without spending too much."
- "1,000 – $1,200"
- "1,600 or less. My annual income is $70k, so this laptop needs to be within my budget. If it is more expensive than that, I will look for something more affordable."
- "1,500 is the maximum price I would be willing to pay for this Surface laptop. I understand that this is a high-end laptop with excellent specifications and I am confident that I can acquire it for that price or lower. I will check online to compare prices and see if I can find a better deal."

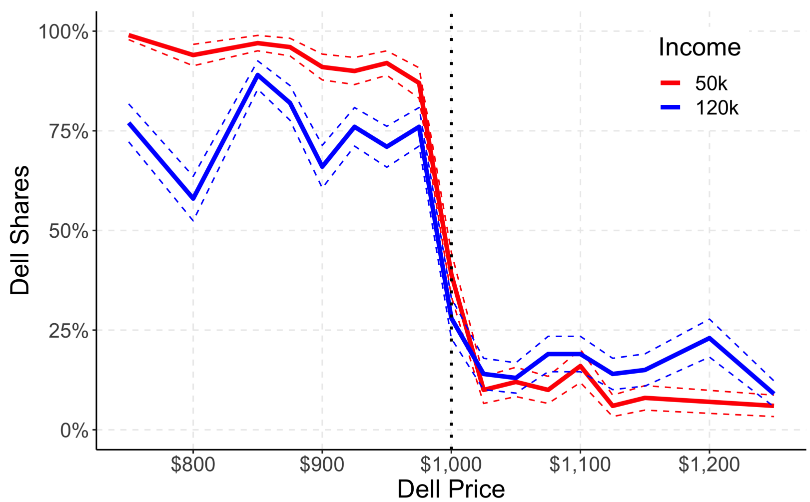
D Robustness: Study 2 – Impact of Income on Demand

In Section 3.1.2, we presented the results of a comparison between Surface Laptop 3 and MacBook Air. Here, we provide two alternative specifications.

Figure 8: Impact of Income on Demand



(a) Surface Laptop 4 (vs. Lenovo Thinkpad)



(b) Dell XPS (vs. HP Spectre X360)