Using GPT for Market Research*

James Brand† Ayelet Israeli‡ Donald Ngwe†

March 21, 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 marketing researchers and practitioners. In contrast to prior work, we focus on the distributional nature of LLM responses. We offer two sets of results. First, we show that the Generative Pre-trained Transformer 3 (GPT-3) model, 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 are of reasonable 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 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 donaldngwe@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 and accurately. LLMs power several well-known AI-augmented solutions for coding (Github Copilot) and search (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, "GPT-3" or "GPT" henceforth) can be used as a tool for market research. Existing tools for market research, such as conjoint studies, focus groups, and proprietary data sets, can be expensive and static. If LLMs can generate responses that are consistent with existing economic and marketing research, then they may also be able to serve as a fast and low-cost method of providing the same information typically generated by conjoint studies and other costly 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 strategy prior to the launch of a new product, and then iteratively querying the LLM over time to evaluate product-market fit and modify the marketing strategy.

We emphasize that, *ex ante*, it is unclear what we should expect to learn from GPT's responses 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 include mentions of 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? Moreover, even if GPT can generate reasonable responses to each isolated question, there is no guarantee that its responses *across* different questions will be internally consistent in the ways we expect consumers to be. Evaluating these sorts of fundamentals is key to understanding the potential value of GPT and other LLMs for almost any marketing analysis, and is the starting point of our paper.

A large literature documents the differences between customer surveys, which elicit stated preferences over bundles of goods, and real world demand data, which informs us of customer preferences 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. 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, empirical features of consumer demand by running four experiments. In each experiment, we provide GPT with a series of prompts, varying one key feature of the choice setting or of customer attributes (e.g., prices, prior purchases, income). These experiments 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. We find in each of these experiments that our estimates of WTP are of reasonable magnitudes, and 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. (forthcoming). 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. Where a survey of real customers may cost many thousands of dollars and take weeks or months to implement, each of our experiments takes less than a couple of hours to run and many are done in minutes.

Although these results are promising, they are only the beginning of research in this area, and more work is needed to identify best practices for learning customer preferences from LLMs. As such, we provide some guidance on the limitations and issues we found while conducting our experiments in Section 4. 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 which were whole dollar amounts, "a single price in dollars and cents" or just "a single price" eliminated this behavior. We also found that, like human survey participants, GPT exhibits response order bias and is much more likely to choose the first response in a binary choice than the second. Thus, although we find success from a GPT-based conjoint in Section 3.2.3 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 mentions the relationship between stated and revealed preferences and concludes that the corpus on which LLMs are trained make the better comparison likely to be revealed preferences, focusing on classic experiments from behavioral economics. He also compares GPT to a random number generator, related to our use of it, focusing 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. 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. Timoshenko and Hauser (2019) use an ML approach on user-generated content (UGC) to identify customer needs for new product development, and Burnap et al. (forthcoming) use generative AI images to predict customers' evaluations to new product design. We contribute to this stream of the literature by further demonstrating how generative AI can provide information on consumer preferences and perhaps simulate market scenarios.

The broader literature on generative AI has identified several means by which AI can improve productivity. Peng et al. (2023) show that Github Copilot, an AI pair programmer, improved the productivity of programmers by 55% in a controlled experiment. In another online experiment, Noy and Zhang (2023) show that college-educated workers to whom ChatGPT, a version of GPT-3 optimized for dialogue, was exposed completed a writing task in 0.8 standard deviations less time and with quality of output rated 0.4 standard deviations higher. 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 level.

2 Querying GPT for Market Research

In this paper we focus on GPT-3 (Generative Pre-trained Transformer 3, "GPT" henceforth) 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 (the latter used for our experiments here) and to

receive many different responses at once for each prompt. One key difference between our study and existing work to date is our focus on the distributional nature of LLM responses. When a worker is using an LLM to accelerate or improve their own output, the predominant feature of the LLM is its ability to reliable 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. It is unlikely that this form of interaction with LLMs is ideal for understanding customer preferences, which is the focus of our work here.

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 bundle 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. Toward this end, we set the "temperature" on GPT to its maximum value (1) for all experiments in an attempt to maximize the variation across responses.[1] Our prompting approach then proceeds as follows. In each experiment, we prompt GPT to fill in the responses of a survey question as if it were a customer, shopping in a category of interest, and were randomly selected to participate in a survey. We describe any relevant features of the customer (e.g., annual income), offer one or two products for this customer to consider purchasing, and then remind the customer that they can always choose not to purchase either of the available goods. We then ask GPT to fill in the response of this chosen customer by ending our prompt with "Customer:".[2] We submit each of these prompts to GPT hundreds of times and aggregate the responses to construct our outcomes of interest.

We take this approach for two reasons. First, in early testing, we attempted other approaches to querying demand-relevant information and found only limited success. In one example, when we asked GPT to provide the likely market shares of two goods, we found it more likely to oscillate between extreme answers than intuition would suggest. Often, GPT would give one product or the other a 100% market share. We also found that asking GPT to fill in multiple responses resulting from a hypothetical survey of customers often generated bunching around one option or another, even when the number of queried responses was large enough to make this unlikely to arise from random sampling. These findings were preliminary, and

[1] 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.

[2] We include examples of the prompts we used in the Appendix.

better prompt engineering may improve these directions for preference elicitation, but these results were in contrast to our near-immediate success with the querying approach we take in this paper.

The second reason for our approach is that we hypothesize that the information in GPT's training set which is most relevant to demand is some combination of product reviews, either formally on merchant websites or on informal internet forums, product descriptions, and any discussions of the underlying product attributes. These source materials are most often written in the first person, meaning that GPT is most likely to be able to fill in reasonable responses to survey questions from the perspective of the customer.

3 Results

We present results from experiments designed to assess the usefulness of GPT as a tool for market research. In our first set of experiments, we study whether GPT's responses broadly align with predictions from economic theory. In our second set of experiments, we compare GPT's responses to benchmarks that are representative of established market research tools: conjoint analysis and demand estimation.

3.1 Testing Predictions from Economic Theory

In our first set of experiments, we enumerate four fundamental demand patterns that are both predicted by economic theory and widely documented in the economics and marketing literatures. We show that, by and large, GPT's responses align with expected consumer behavior. For each of these experiments, we prompt GPT as described in the previous section, varying one attribute of the offered choice at a time and aggregating hundreds of responses for each prompt. Further details, including the specific prompts we use, are provided in Appendix Section A.

3.1.1 Experiment 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 experiments by establishing how GPT responds to price changes for a single good, holding everything else constant.

We conduct three separate exercises to document the shape of the demand curve from GPT

responses, the results of which are shown in Figure 1.[3] 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. In this simple scenario we do find that the demand curve is downward-sloping. 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 magnitude of the decrease in demand generated may be unrealistically small.

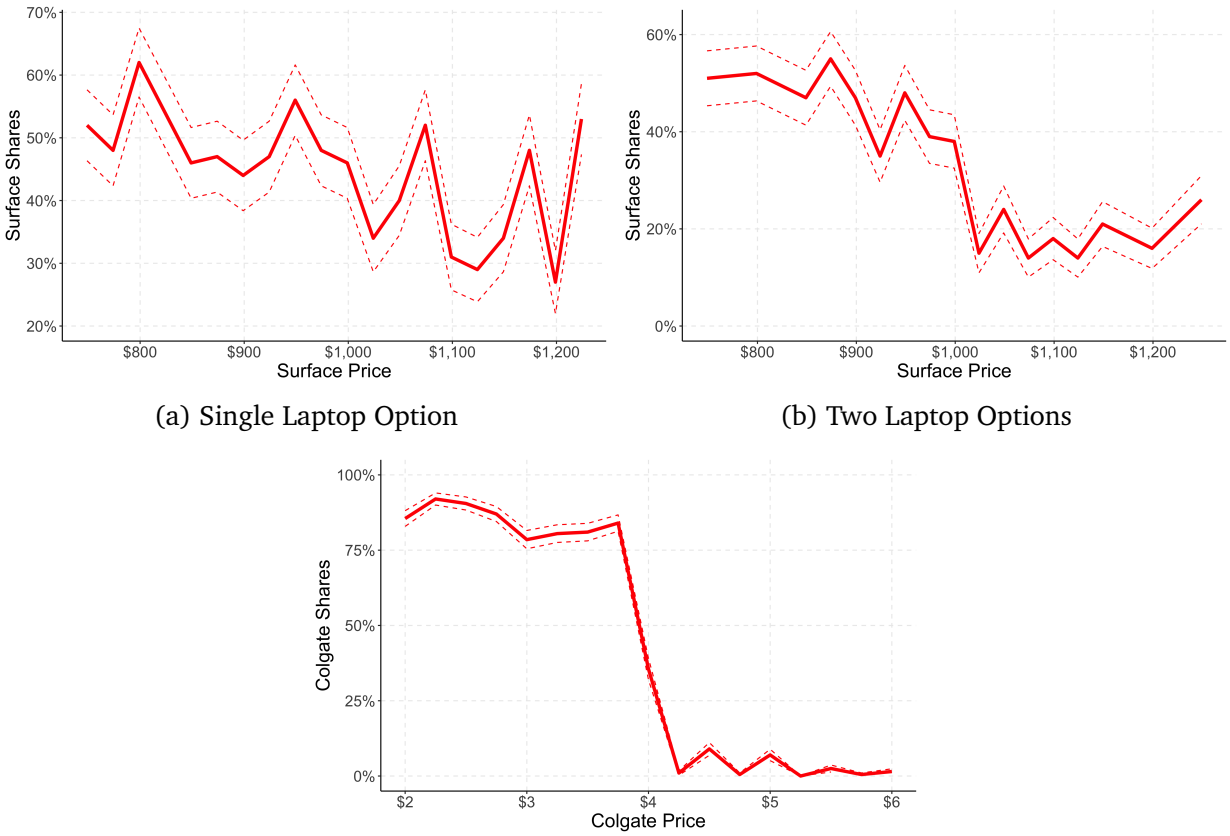
We note as well that this and the succeeding demand curves implied by GPT are generally not *monotonically* downward-sloping. We interpret this characteristic as a feature and not a bug: much work in economics and marketing has documented several possible relevant factors, including preferences for round numbers, prices as a signal of quality, and left-digit effects (e.g., 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.

For our second and third exercises, we 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) at a fixed price of $999 while varying the price of the Surface 3 laptop. Here we find a much more steeply sloped demand curve. Notably, although the demand curve is downward sloping over much of the domain, we see a particularly sharp drop in demand for the Surface option around the price of the reference good.

For our final exercise 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 from this final figure. First, the demand curve continues to be broadly downward-sloping. Second, demand for the focal good in this setting appears to decline much more quickly than in the previous experiment. When the focal good is even marginally more expensive than the reference good, demand for the former drops nearly to 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 is consistent with both higher levels of perceived horizontal differentiation for laptops than for toothpaste and with prospect theory (which suggests that customers should be more loss averse to smaller changes in prices).

[3] 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.

Figure 1: Downward Sloping Demand Curve



(a) Single Laptop Option

(b) Two Laptop Options

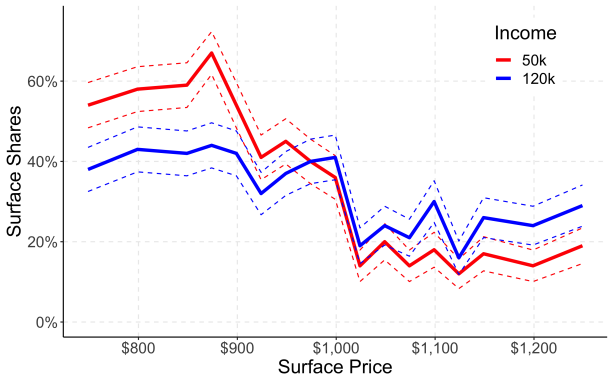
(c) Two Toothpaste Options

3.1.2 Experiment 2: Declining marginal value of wealth

After verifying that demand curves estimated from GPT survey responses are generally downward sloping, our next experiment focuses on the shadow value of wealth. Economic theory and empirical work suggest that wealthy customers are less sensitive to changes in price than poorer 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, in an attempt to generate behavior similar to the median household.[4] In Figure 2, we submit the exact same set of prices and products to GPT that we did in order to generate Figure 1b, but change this stated income, first to $50,000 and then to $120,000. Theory predicts that the demand curve should be more shallow when GPT is prompted with a higher income, and this is indeed what we see. When the focal good's price exceeds that of the reference good by $250, GPT customers with $50,000 of income reduce their demand for the focal good to roughly 15% while GPT customers with $120,000 of

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

Figure 2: Income Prompting: Diminishing Marginal Utility of Wealth



income continue to choose the focal good about 25% of the time. In Appendix Section B we show that this result is robust to choosing different set of laptops and reference prices.

3.1.3 Experiment 3: Diminishing marginal utility

In our third set of experiments we examine whether GPT's responses reflect a diminishing marginal utility of consumption. The law of diminishing marginal utility is a central tenet of consumer theory and is at the heart of the explanation for numerous economic phenomena. Within market research, the extent to which marginal utility diminishes is useful for setting quantity discounts, demand forecasting, and inventory management.

For the first of these experiments, we modify our prompt with a statement indicating that the randomly selected customer being interviewed has already purchased the good in the past and has a given number X units of the good at home, varying the prompted value X from 0 to 1,000.[5] In this exercise, we focus on yogurt, which is often purchased in packs of 4-12 6 oz. cups, as a realistic setting in which a customer may both have a stock of a supermarket good but also consider purchasing another. For each value of X, we then ask GPT to provide the customer's willingness to pay for an additional unit of yogurt at the store.

Figure 7a 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. We find a sharp decline in the mean reported WTP from $X = 0$ to $X = 1$, followed by a series of very similar values for all X between 1 and 10. The final few columns of our plot focus on much larger values of X, ranging from 20 all the way up to 1,000. It is in these columns where diminishing marginal utility most clearly sets in. A customer with 20 units of yogurt at home already is willing to pay approximately 15% less at the median than a customer with 10 or fewer pre-purchased units.

[5] Specifically, we used whole unit values between 0–10, and also 20, 50, 100, and 1000.

The average WTP is also decreasing 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 what we expected.

Because in the yogurt purchase scenario GPT may infer additional information (e.g., stockpiling, bundling, quantity discounts) from the fact that the customer has X units at home, we moved to replicate our diminishing marginal utility findings in an immediate consumption context — beverage consumption at a restaurant. In this exercise, we focused on glasses of beverages (soda and wine), and asked a "random restaurant goer" how much they would be willing to pay for an additional glass after having ordered and consumed X glasses. Figures 3b and 3c present the results. Unlike the yogurt scenario, we do not find evidence of diminishing marginal returns in these scenarios.

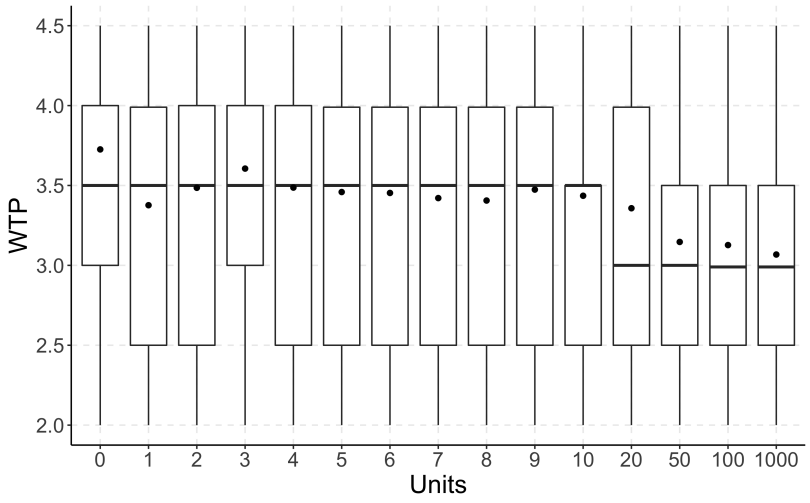
As with our experience above with GPT's responses yielding non-monotone demand curves, we recognize that isolating specific relationships (e.g., utility of money or diminishing marginal utility) using consumer surveys is challenging and not the usual objective of survey instruments. In this case, prompting 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 simply does not easily simulate diminishing marginal utility, it is also possible that in practice several other factors impact the relationship between consumption levels and willingness to pay for the marginal unit.

We do note that the method by which we have tested diminishing marginal utility is one that is uniquely suited to LLMs. In many customer surveys, 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. In this experiment, we instead simply ask the GPT customer to provide their willingness to pay directly. We will return to the question of the realism of these stated values in Section 3.2.

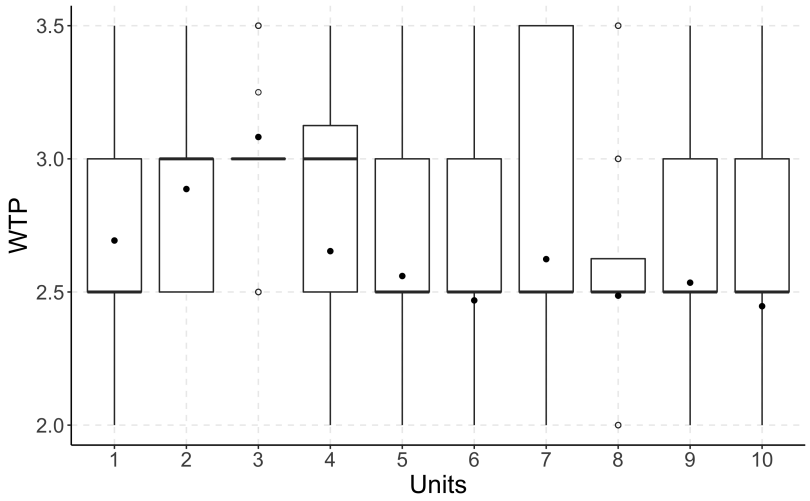
3.1.4 Experiment 4: State dependence

Our final experiment is designed to test whether GPT's responses are consistent with state dependence. There is a significant body of work in industrial organization and marketing which has studied the magnitude and causes of serial correlation in customers' choices in a variety of contexts. In health insurance markets, for example, there are concerns about inertia as a reason that customer repeatedly choose the same plan even when the plan's attributes change significantly (e.g., Handel, 2013; Pakes et al., 2021). There are also a number of papers discussing the tools and data necessary to distinguish between various forms of state

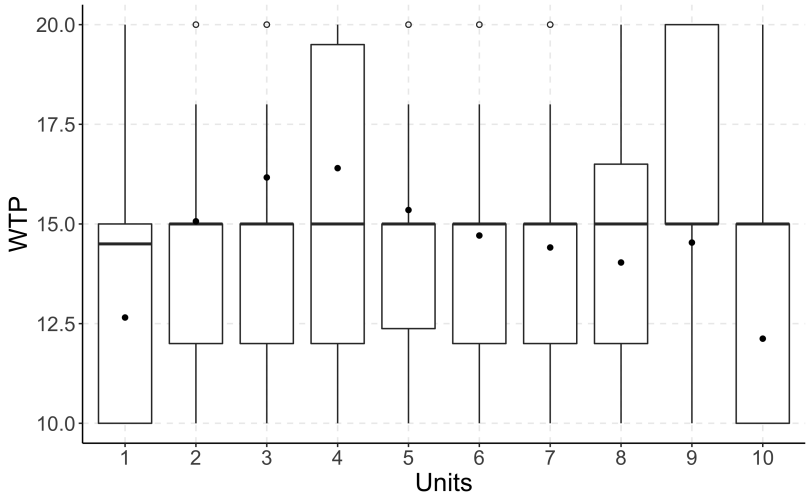
Figure 3: Diminishing Marginal Utility of Consumption



(a) Yogurt



(b) Glass of Soda at a Restaurant



(c) Glass of Wine at a Restaurant

dependent choice and preference heterogeneity in the choice of consumer packaged goods (Dubé et al., 2010; Levine & Seiler, 2022).

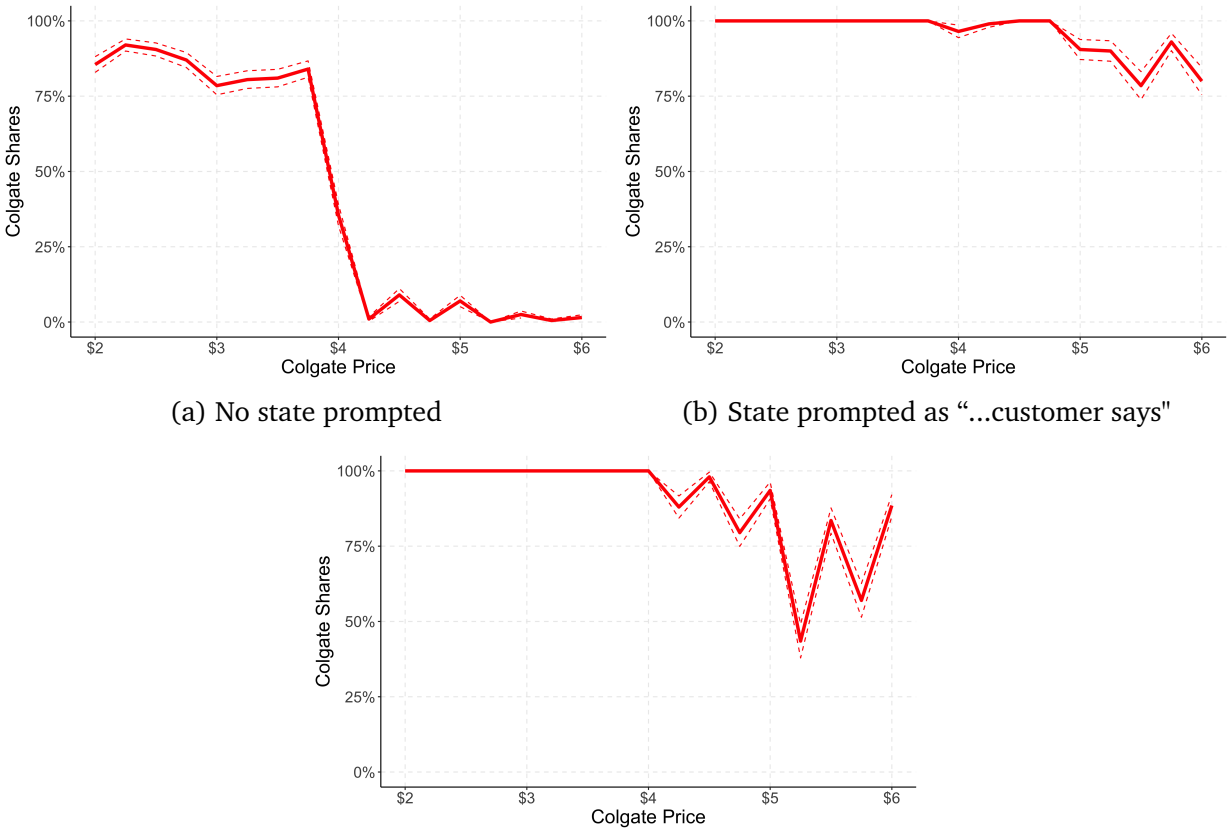
For this experiment, we conduct the same experiment as in Figure 1c (choice between two brands of toothpaste, with Colgate being the focal brand), but now include a phrase in the prompt indicating the good that the customer purchased previously. In Figure 4a, we copy the original result with no state-related prompting for reference. Then, in Figure 4b, 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. For any price such that the focal good (Colgate) is cheaper than the reference good, we find that all responses choose Colgate. Only once Colgate is more expensive than the alternative do we see the demand curve begin to slope downward again. For our final exercise, as a way of highlighting the nuances of prompting GPT for our experiments, in Figure 4c we plot the demand curve generated by a prompt which 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 which, while less clearly monotonic, has a significantly steeper slope than in Figure 4b. 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.

3.2 Contextualizing GPT Responses

The previous section demonstrates that GPT's answers 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. Moreover, for results like those in Section 3.1.3 which do target levels, it is possible that what is being returned by GPT is not any coherent willingness-to-pay metric, but rather an attempt by the LLM to match the distribution of prices listed on websites selling the goods in the prompt.

We address these concerns using three different experiments. First, we demonstrate that the distribution of willingness-to-pay for products, produced by GPT via the direct solicitation approach used in the previous section, generates reasonable values for multiple categories of goods. Second, we show that we can also back out GPT's willingness-to-pay for product attributes, taking flouride in toothpaste as an example. Willingness-to-pay for flouride is much less likely to be stated directly on product pages or in reviews than is the toothpaste's price, but GPT is still able to generate reasonable estimates WTP estimates from both direct and indirect

Figure 4: State Dependence: Previous Colgate Purchase



(a) No state prompted

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

(c) State prompted as fact

elicitation approaches. Finally, we recover preferences via a conjoint-style survey of GPT, and demonstrate that our results match those in Fong et al. (forthcoming), who conduct surveys of real individuals, on multiple dimensions.

3.2.1 Experiment 1: Recovering realistic WTP For products

In this section we aim to understand whether asking GPT directly for willingness to pay (WTP) for certain products provides a meaningful/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). We provide the same specifications we used in earlier prompts, but do not include any price. The median implied WTP for the Surface Laptop 3 is $1,000, similar to its market price.

Next, we use more general descriptions to elicit WTP for a good, rather than WTP for a particular brand. Recall also from section 3.1.3 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 the restaurant scenario with a median of $2 (see Figure 5d).

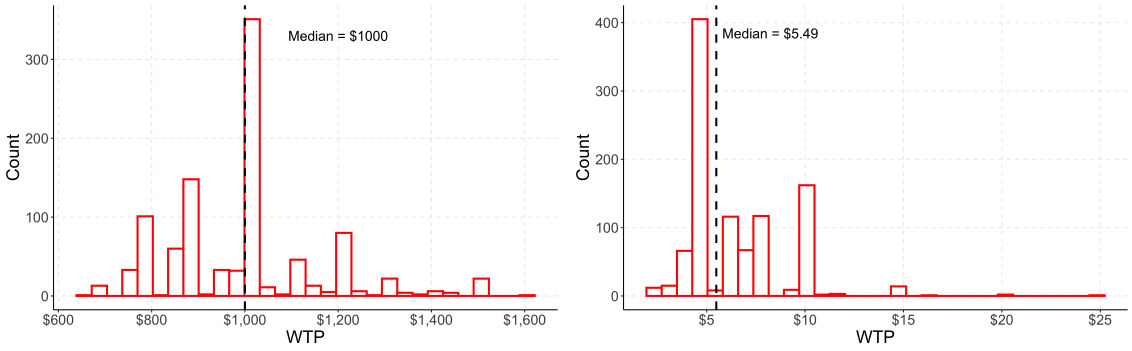
3.2.2 Experiment 2: Recovering realistic WTP For attributes

After demonstrating that asking GPT directly for WTP provides a realistic distribution of values for a variety of goods, we move on to examine whether we can recover estimates of WTP for attributes from GPT's responses. In this section, we use toothpaste as the product and fluoride as the attribute (borrowing from Fong et al. (forthcoming), which we will also use in Section 3.2.3).

We utilize two different strategies in this section: a direct elicitation approach, and an indirect elicitation approach. For the direct elicitation approach, we offer two identical toothpastes that differ only on the existence of fluoride and ask GPT how much more it would be willing to pay for the option with fluoride over the option without fluoride. Figure 6a provides the distribution of WTP for fluoride using this approach.

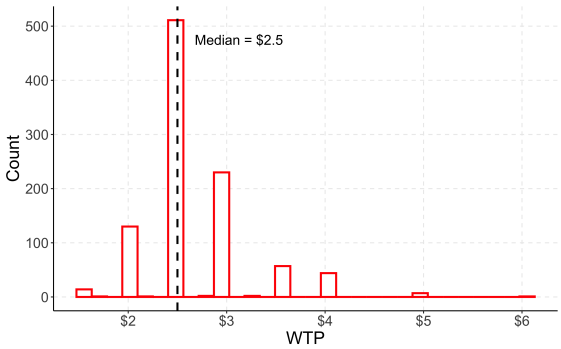
The indirect elicitation approach consists of two steps. First, we use GPT to estimate the demand for Colgate whitening toothpaste without fluoride using a similar paradigm that we used in Section 3.1.1 to generate a demand curve (this time, the focal good did not have flouride, but the reference good priced at $4 did have flouride). Then, we compare the demand curve for the toothpaste with fluoride (from Figure 1c) with the demand curve for the toothpaste without fluoride (see Figure 6b 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 demand for toothpaste with flouride at p' is equal to the demand for flourideless toothpaste at p. Our WTP measure is then $p' - p$, which is a function of p and which amounts to taking horizontal differences between the demand curves in Figure 6b. This approach suggests, for example, that when the price of Colgate without fluoride is $3, the WTP for fluoride is $1.19. We note that this is similar to the median WTP of $1 generated by the direct elicitation approach in Figure 6a.

Figure 5: Willingness to Pay for Products, Direct Elicitation

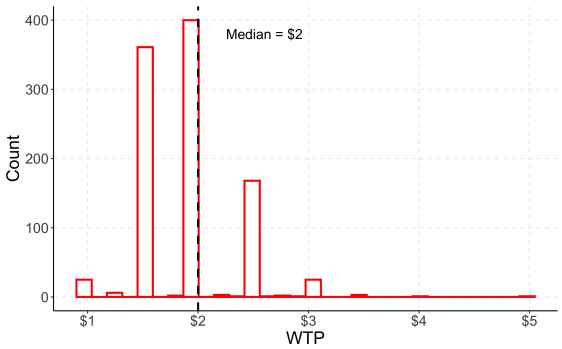


(a) Surface Laptop 3

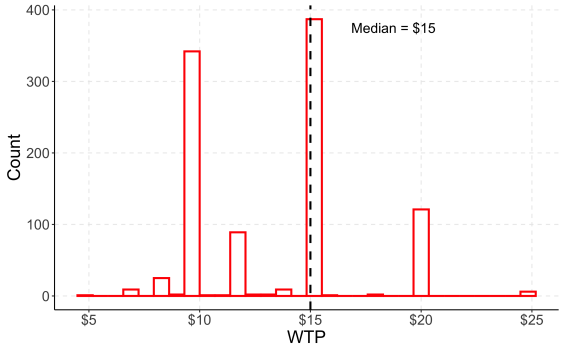
(b) Whitening Toothpaste



(c) “Glass of soda (at restaurant)"



(d) “Can of soda (at supermarket)"

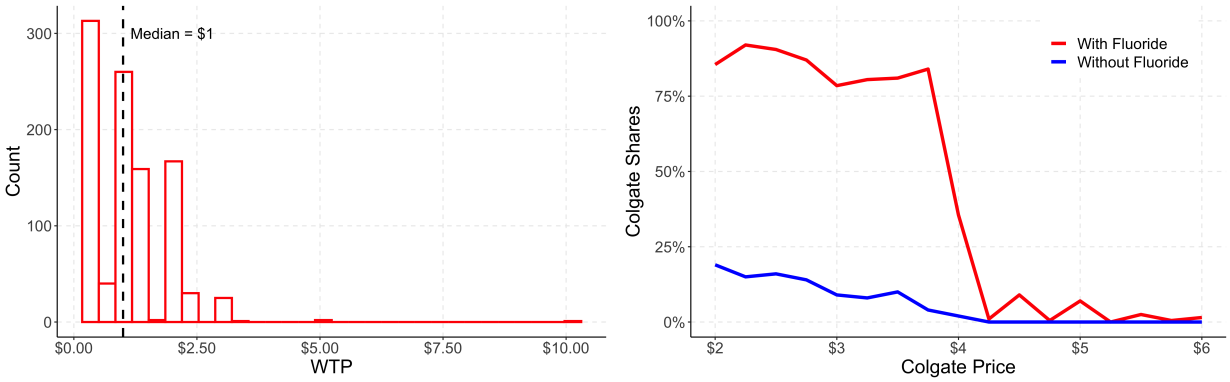


(e) “Glass of wine (at restaurant)"

3.2.3 Experiment 3: Recovering preferences via conjoint

For our final experiment, we focus on recovering preferences using the conjoint analysis paradigm. Conjoint is widely accepted in industry and academia for this purpose, and has been shown to be able to uncover customer preference for different product attributes jointly (see Green and Rao (1971), Green and Srinivasan (1978), and Green and Srinivasan (1990) for a review).

Figure 6: Willingness to Pay for Attributes



(a) Fluoride, Direct Elicitation

(b) Fluoride, Demand curve

In this section, we attempt to evaluate GPT on two dimensions. First, we treat the responses from GPT as if they were from randomly chosen consumers and test whether the effect of prices and non-price attributes on choice probabilities are consistent with economic predictions. Although we have shown above that GPT's demand is decreasing a product's own price when the other price is fixed, our experiments here test whether the same holds true (on average) in more complicated setups in which the prices and other attributes of both goods can vary across prompts. Second, we use the queried responses to estimate a multinomial logit model similar to the kind that practitioners use to estimate preferences in standard conjoint analysis, in order to evaluate the realism of model-based estimates of WTP.

For this section we chose to focus on choices of toothpaste. We were inspired by Fong et al. (forthcoming), who recently ran a thorough conjoint study and confirmed that their experimental estimates are consistent with market outcomes. We focus on the two brands we used for our earlier analyses, Colgate and Crest. Similar to Fong et al. (forthcoming), we use 3 levels of prices ($0.99, $1.99, $2.99), and 2 fluoride options (with, without). When designing conjoint studies, the typical methodology, for practicality and soundness purposes, is to derive a few choice sets which are orthogonal across configurations and that balance attributes across choices. Then, study participants are typically presented with 10–15 scenarios comparing 2–3 products (as well as a no-choice option). Because we are not limited by humans' time or ability to process complex information, we chose to create the full set of options for each brand: three price levels for each of two fluoride options, yielding a total of 36 configurations. We collect 300 responses for each configuration, for a total of 10,800 responses. Overall, 2,300 responses chose Crest, 2,468 chose Colgate, and the remainder opted out from making a choice.

Our first set of results are shown in Table 1. We first present simple regressions and separate es-

timates for each brand's price and fluoride attribute using ordinary least squares. Reassuringly, the estimates are in the expected signs: when Crest (Colgate) is priced higher, the likelihood of choosing Colgate (Crest) is 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: 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
Note:		*p<0.1; **p<0.05; ***p<0.01

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. 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 section and are quite similar to the estimates in Fong et al. (forthcoming), who conduct a real-world conjoint to estimate customer preferences for toothpaste and estimate the WTP for fluoride to be $3.27. Table 2 includes both the multinomial logit results in column 1 and a random coefficient model in column 2. The results are consistent across both estimation methods (in the random coefficient model our WTP estimate is $3.30).

Table 2: Conjoint Results

	(1)	(2)
Price	−0.484*** (0.021)	−0.504*** (0.034)
Fluoride	1.647*** (0.037)	1.662*** (0.044)
Colgate Brand Dummy	−0.801*** (0.051)	−0.778*** (0.060)
Crest Brand Dummy	−0.491*** (0.050)	−0.457*** (0.063)
σ Price		0.155** (0.067)
σ Fluoride		1.049*** (0.149)
Observations	10,800	10,800
Note:		*p<0.1; **p<0.05; ***p<0.01

4 Guidelines and Limitations in Querying GPT

While running the experiments in this paper, we have identified 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. We offer a selection of examples here, while recognizing that these are a small representation of a full set of guidelines for using GPT in market research.

Sensitivity to Response Order. We found in early work that, 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%. 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.

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 a 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 understanding customer preferences. Our first set of experiments 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 wealth and state dependence. These two properties in particular are notable because of their complexity. Not only are the demand curves elicited from GPT downward sloping, but when we are querying multiple demand curves under different scenarios (i.e, incomes, historical purchase behavior), the relationship between different demand curves is similarly coherent. This is a surprising, but essential feature for these types of systems to have promise as tools for marketing researchers and practitioners.

Our second set of results demonstrates that the empirical quantities derived from GPT's responses are realistic and consistent with values obtained from existing research. We begin by directly eliciting WTP for multiple categories of goods and find that the reported prices are of reasonable magnitudes, while noting that such direct elicitation of WTP from human subjects is known to suffer from many shortcomings (Wertenbroch & Skiera, 2002). We then demonstrate that GPT can generate more complicated objects of interest to marketing researchers. GPT is able to generate estimates of WTP for fluoride that are close to existing research when given a conjoint-style survey, 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.

Just as there is a substantial literature discussing the best ways to elicit customer preferences for goods, we envision a similar literature will evolve for GPT-based surveys. We have demonstrated significant potential for progress here, but translating these results into practical tools for researchers and businesses will require many iterations. At present, we can easily 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 that may be more realistic than standard approaches to simulated data, given the emergent properties highlighted here and others that are likely to be found in the future.

Second, users may adapt GPT to suit specific contexts by endowing it with “knowledge" of various forms. In our simple exercises we endow GPT with income and prior purchases; one might imagine assigning GPT personas, preferences, and product or market information of increasing complexity. Moreover, researchers may wish to integrate over such inputs according to the empirical distribution of the traits of interest. For example, to generate a nationally representative sample, one could inject incomes into the prompt as we have done here, but draw the prompted incomes from the full national income distribution and average over all incomes to generate aggregate objects of interest. Many other similar directions are apparent and may provide even more realistic estimates than we have found here, and in cases where the researcher has convincing prior knowledge about some moment of the WTP distribution, there may be opportunity to conduct calibration exercises where the prompt design is engineered to match the known moment and other moments for analysis are left free.

Finally, we expect that LLMs’ usefulness for market research will increase in lockstep with the rapid improvement in these models’ sophistication. 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.

To conclude, we 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 apparent slowness to reflect diminishing marginal utility. We expect that there are at least a few more. 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. Further, 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

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. (forthcoming). 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*.

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

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

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.

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 Prompts and experimental 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.

A.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 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: "

A.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.

A.3 Prompts for Section 3.1.3

- 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: "

A.4 Prompts for Section 3.1.4

- 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: "

A.5 Prompts for Section 3.2.1

- Laptop:

  For the supermarket examples:

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

  While shopping, the customer sees a 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: "

A.6 Prompts for Section 3.2.2

- For direct solicitation:

"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:

  "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: "

A.7 Prompts for Section 3.2.3

"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: "

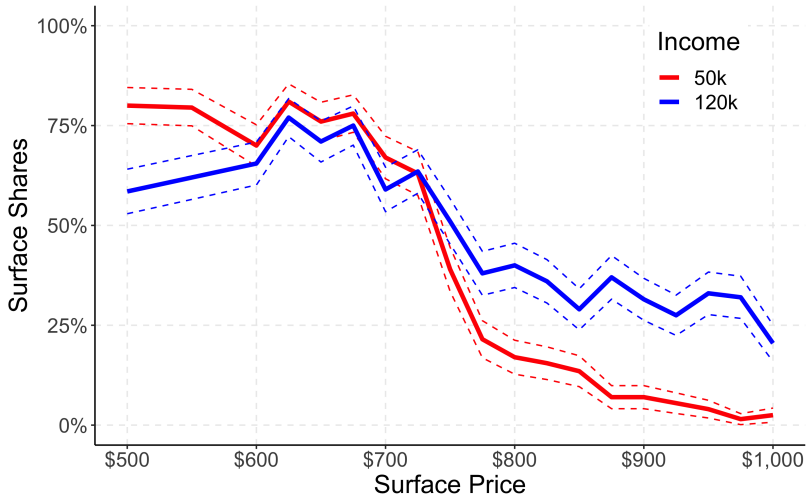
A.8 Number of observations collected

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

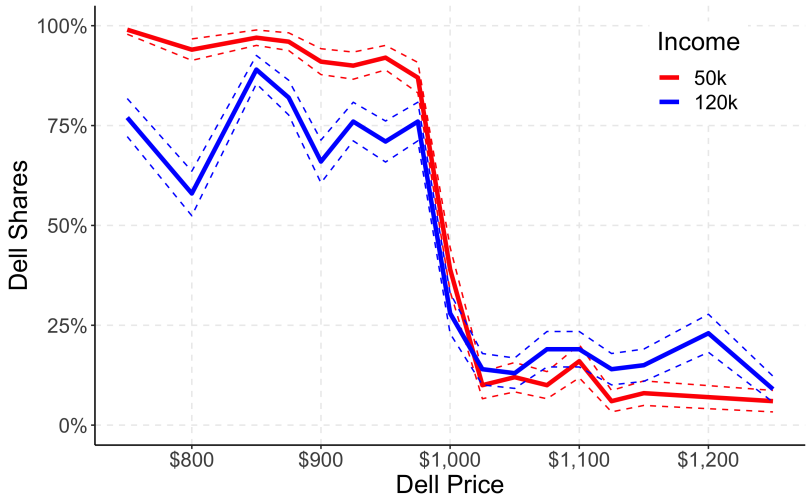
B Robustness: Experiment 2 – Diminishing marginal utility of wealth

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 7: Diminishing Marginal Utility of Wealth



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



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