Using LLMs for Market Research*

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

July 29, 2024

Abstract

Large language models (LLMs) have rapidly gained popularity 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 Turbo (GPT-3.5 Turbo) model to generate dozens of responses to each survey question. We offer two sets of results to illustrate and assess our approach. First, we show that estimates of willingness-to-pay for products and features derived from GPT responses are realistic and comparable to estimates from human studies. Second, we demonstrate a practical method for market researchers to enhance GPT's responses by incorporating previous survey data from similar contexts via fine-tuning. This method improves the alignment of GPT's responses with human responses for existing and, importantly, new product features. We do not find similar improvements in the alignment for new product categories or for differences between customer segments.

*The authors are grateful to Noah Ahmadi and Meng Yang for excellent research assistance. This manuscript was previously circulated with the title "Using GPT for Market Research."

[†]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, Gemini), 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, primarily the Generative Pre-trained Transformer 3.5 Turbo model, "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 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 hypothesize that the responses GPT provides to market research surveys will, in part, reflect the types of responses that the customers in the training data would have given to the same questions. Together, these components would 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 supplementing 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 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 in this context. 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 and 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 LLMs responses to prompts, motivates our investigation into the usefulness of LLMs for market research.

To quantify GPT's efficacy in this setting, our empirical analysis attempts to approximate problems that market researchers might face in practice. In many cases, market research is used to learn something about customers' preferences, either for existing products and features, new features for existing products, or entirely new product categories. In this context, "new" can either refer to products/features for which the firm has little historical information or to innovative features which have never been created before. Moreover, marketers often care both about the preferences of a generic market customer and about the segments of the market they expect to target. With these use cases in mind, we design five surveys of human participants that allow us to isolate the information set of a hypothetical market researcher and to explore the performance of GPT in each case. We begin by studying GPT's ability to simulate consumer preferences in settings where the researcher has no prior information about the product category, and then add multiple levels and forms of information about the products and customers of interest, which come from our human surveys. These studies simulate the steps a market researcher might take to use previous studies of the same or similar product markets and customer segments to supplement GPT.

Overall, the resulting WTP estimates are 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), as well as to additional human surveys we conduct. Furthermore, we find that augmenting GPT with data from human surveys improves not only GPT's ability to respond consistently with those surveys but, remarkably, to match separate human surveys on preferences for new product features. In contrast, applying this approach to different product categories does not yield more aligned results from GPT. Together, our results suggest that GPT may provide an alternative means for marketers to learn about consumer preferences in a fast, low-cost, and iterative manner. While we do not suggest that marketers use GPT to replace humans, our approach may help test out and narrow down new feature ideas before testing them with humans, particularly when marketers already have surveys from the population of interest. 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 hours, and the projected costs are substantially lower than those of human samples.

We have not, in our investigation, been able to have GPT meaningfully reflect heterogeneous preferences across demographic groups. We explore, through several distinct approaches, whether GPT can reflect heterogeneity in WTP across multiple attributes including gender, income, and political affiliation, and conclude that even with our augmentation method, the current state-of-the-art models of GPT better reflect average population WTP and do not provide meaningful across-group estimates.

Although our results are promising overall, research in this area is preliminary, and more work is needed to identify best practices for learning customer preferences from LLMs. In Online Appendix A, we provide some guidance on the limitations of the approach and issues we encountered while conducting our studies.

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 LLMs provide responses to classic behavioral economics experiments 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. He also compares GPT to a random number generator, which is related to our approach. We focus on the distribution of responses rather than a single draw. Subse-

quent work in a variety of fields (e.g., Bisbee et al., 2024; Dominguez-Olmedo et al., 2023; Goli & Singh, 2024) examines the limitations of using LLMs as synthetic survey respondents, highlighting challenges in response quality and reliability, and proposing solutions to mitigate these concerns. We add to this literature by evaluating LLMs' ability to extract customer preferences.

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. (2023) 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. (2024) demonstrate how to use LLMs to construct perceptual maps by querying them about brand similarities, and show that the responses are similar to those of humans. We contribute to this stream of the literature by further demonstrating how widely available generative AI tools may be able extract consumer preferences.

2 Research Design

This section provides an overview of our research design, focusing on how we use GPT as a tool for simulating human responses in market research. We detail our approach to querying GPT, including the selection of the GPT-3.5 Turbo model for our studies, and our methodology for evaluating its performance in reflecting consumer preferences. Additionally, we explain our use of conjoint analysis to estimate willingness-to-pay (WTP) for product attributes, both from human participants and GPT-generated responses. The section also explores how we simulate demographic heterogeneity and incorporate fine-tuning with existing survey data to enhance GPT's responses.

2.1 GPT

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 common programming languages[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

[2]For the majority of our studies, which we ran between April and July 2024, we used GPT model "gpt-3.5-turbo-0125." We also used the "text-davinci-003" for a few of our pilot studies, which ran in March 2023. We typically access the API using Python. See Appendix A for a code sample.

nature of LLM responses. Knowledge workers, for example, may use an LLM to accelerate or improve their 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.

We use the GPT-3.5 Turbo (more formally, OpenAI's "gpt-3.5-turbo-0125" model) as the main model for our studies for two reasons. First, it was the latest model that was recommended model by OpenAI for fine tuning, a feature we were interested in exploring, and we wanted to ensure all our results rely on the same model.[3] Second, GPT-3.5 Turbo was OpenAI's most cost-effective advanced model.[4] To examine the robustness of our methodology to choice of GPT model, we also replicated one of our studies using a variety of models from OpenAI, Meta, and Anthropic in Section 3.1.3.

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). Our hypothesis is that, when we induce GPT (or other LLMs) to provide a choice 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 humans it aims to represent.[5] Hence, our approach to performing market research using GPT is to treat the model as a tool for predicting real consumer responses to surveys 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

[3] While OpenAI now has a wait-list for experimental fine-tuning with new models, the recommended and widely available model for fine-tuning is still GPT-3.5 Turbo.

[4] For GPT-3.5 Turbo the costs for API calls are (for 1M tokens): Input tokens: $0.50, Output tokens $1.50, whereas GPT-4o is: $5, and $15, respectively. Fine tuning token costs are significantly higher: Input: $3, Output: $6, whereas fine tuning a model costs only $8 for 1M training tokens.

[5] "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.

including the well-known Milgram 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

At a high level, our research design considers the problem of a market researcher seeking to estimate consumer preferences when they are unable to conduct new human surveys that directly address the customers or products of interest but may have related data from previous surveys. We consider two ways in which GPT can help the researcher in this setting. First, they may survey GPT directly and use that survey data heuristically as a supplement to other human studies. Second, they may wish to supplement GPT's knowledge with available data and to focus GPT's responses on a desired target population. In these cases, they may wish to find ways to incorporate surveys run by their firm in the past, in settings that are distinct but similar to their current focus. In this section, we outline our strategy for evaluating GPT for both of these use cases.

We focus on the specific problem of estimating WTP for product attributes, a task that often falls to market researchers. To calculate WTP for the attributes we study, we recover preferences for attributes and for prices 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 Srinivasan, 1978 and Green and Srinivasan, 1990 for a review). Another reason we choose to focus on WTP for product attributes rather than for products is that compared to WTP for products, WTPs for attributes are much less likely to to be stated directly on product pages or in reviews and to thereby appear in GPT's corpus.[6]

[6]In Online Appendix B we demonstrate that the distribution of WTP for products generates realistic values for multiple categories of goods. We note the concern that GPT's output reflects the distribution of listed prices rather

We studied multiple approaches to soliciting preferences and determined that conjoint-style questions were most likely to elicit useful results from GPT. Direct solicitation of WTP for product attributes is not the typical method for human-based market research, as humans' ability to quantify these measures is limited. Here, we are using GPT as a simulator of human responses, and therefore we anticipated that the workhorse indirect elicitation methods used on humans are also most likely to be successful with GPT. Nevertheless, prior to focusing on conjoint analysis we tested GPT's ability to provide WTP for attributes directly and through a relatively simple indirect method described in Online Appendix C. Overall, we concluded that these approaches are inferior to conjoint in eliciting preferences from GPT, and therefore we use conjoint for the remainder of the paper.

To evaluate GPT's ability to provide meaningful and realistic WTP, we collect responses to (nearly) identical questions from both GPT and humans. We compare these responses to each other and proceed to examine GPT's ability to reflect consumer heterogeneity. Finally, we evaluate whether and when using existing human surveys to supplement GPT may improve its outputs.

2.3.1 Baseline Human Studies

In order to test the capabilities of GPT in market research applications, we require a reasonable baseline of responses to which we can compare GPT's output. While traditional conjoint surveys have well-known shortcomings in reflecting actual consumer preferences, we rely on human surveys as the best available benchmark, as well as the most relevant to marketers. We use two baselines for customer preferences in lieu of "ground truth." First, as a pilot study, we use existing research by Fong et al. (2023), who conduct a real-world conjoint in late 2020 to estimate WTP for household consumer goods and confirmed that their estimates are consistent with market outcomes, as our baseline for comparison. We ran GPT studies in the spirit of Fong et al. (2023) in March 2023 using the most recent GPT model at that time (text-davinci-003). Importantly, that model's training corpus could not include the results of the Fong et al. (2023) studies given the training cutoff time.

Second, after the pilot study's initial encouraging results, we collected a new set of human conjoint surveys to broaden our investigation, estimate preferences for sub-groups of customers, study new product features, and fine-tune GPT. Toward this end, we began by running five new human studies covering three categories of consumer products: toothpaste, laptops, and tablets. Three of our studies (1A, 2A, 2C) include product features that already exist and are

than consumer WTP, which further motivates our focus on WTP for product features.

thus likely to be in GPT's training corpus, and two studies (1B, 2B) that include new, hypothetical product features. While we use all of the studies to evaluate GPT's baseline responses, we use the latter studies as a human benchmark to evaluate GPT's performance for new features, with and without fine-tuning, as we explain in Section 2.3.4.

All of our studies are outlined in Table 1, and the specific attribute levels used in our studies are outlined in Table 2. In each of these studies, through Prolific we recruited 300 participants 18 or older who reside in the US. Participants responded to 12 choice tasks in which they were asked first to choose their preferred option between two product configurations and then to indicate whether they are interested in purchasing their selection or prefer not to purchase any item at this time.[7] Study materials are available in Online Appendix D.

Table 1: List of Conjoint Studies

Study	Product category	Product attributes	Comparison human sample
Pilot study A	Toothpaste	Fluoride	401 participants, Fong et al. (2023)
Pilot study B	Deodorant	Aluminum	155 participants, Fong et al. (2023)
Study 1A	Toothpaste	Fluoride, conventional flavors	300 participants, this paper
Study 1B	Toothpaste	Study 1A + unusual flavors	302 participants, this paper
Study 2A	Laptop	Storage, RAM, Screen size	302 participants, this paper
Study 2B	Laptop	Study 2A + Built-in Projector	300 participants, this paper
Study 2C	Tablet	Storage, RAM, Screen size	300 participants, this paper

Notes: All studies include also brands and price attributes. All human studies were implemented using Sawtooth Discover software and distributed via Prolific. The human studies in Fong et al. (2023) were conducted in September 2020, and the human studies in this paper were conducted in April 2024.

Table 2: Conjoint Design Attribute Levels

Attribute	Toothpaste	Deodorant	Laptop	Tablet
Brand	Colgate/Crest	Dove/Speed Stick	Microsoft Surface Laptop/ Apple Macbook Air	Microsoft Surface/ Apple iPad
Price	$0.99/$1.99/$2.99	$1.99/$2.99/$3.99	$1,000/$1,200/$1,400	$450/$550/$650
Ingredient	Fluoride (Yes/No)	Aluminum (Yes/No)		
Flavor	Mint/Cinnamon/ Strawberry			
Storage			256GB/512GB/1T	64GB/128GB/256GB
RAM			8GB/16GB/32GB	4GB/8GB
Size			13 in/15 in	8 in/10 in
New features	Pancake/Cucumber flavors		Built-in projector (Yes/No)	

Notes: The pilot studies included only brand, price, and ingredient. The brands and prices for toothpaste and deodorant were taken from Fong et al. (2023), and the laptop and tablet attributes were selected based on products available at the time the human studies were conducted (April 2024). New features were included only in studies 1B and 2B.

[7] We borrow the design of first collecting preferences and only then purchase intent from Moshary et al. (2023).

2.3.2 Baseline GPT Preferences

As in 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. For each set of goods we consider, we query GPT dozens of times as our goal is for GPT to generate a distribution of responses rather than produce a single one. To this end, we set the "temperature" on GPT to its default value (1.0)[8] for all studies in an attempt to maximize variation across responses.[9]

Our prompting approach then proceeds as follows. In each study, we prompt GPT to fill in the responses to survey questions as if it were a customer shopping in a category of interest who was randomly selected to participate in a survey. We may also indicate customer attributes (e.g., annual income). We offer two products for this customer to consider purchasing, and remind the customer that they can always choose not to make a purchase. We record and parse each free form text response. We submit each of these prompts to GPT dozens of times and aggregate the responses to construct our measures of interest. In crafting the prompts for our querying approach, our main goal was to explore and demonstrate ways to use GPT as a market research tool rather than to optimize prompts to achieve a particularly level of accuracy or reliability. We discuss key learnings about GPT prompts in Online Appendix A, and leave prompt engineering and optimization for future work.

Typical conjoint studies require generating choice sets that are orthogonal across configurations and balancing attributes across choices. Study participants are then presented with 10–15 scenarios comparing 2–3 products from the full set of configurations. Because, unlike humans, GPT is not constrained by time or the ability to perform complex comparisons, we chose to generate the full set of options for each set of attributes and create a full-factorial design for each of our conjoint studies. Using this approach, we run GPT-based conjoint studies for toothpaste, laptops, and tablets to correspond with our human studies (see Tables 1 and 2 for details). The exact prompts we use for all of our studies and other design details are available in Online Appendix E. A sample prompt is provided in Appendix B, and sample responses are available in Online Appendix F.

We designed our GPT studies to be as similar as possible to our human studies. In our pilot studies, because we didn't have individual-level human data against which to compare GPT

[8]The possible temperature values are 0–2, where the maximum is 2.0, which generates "creative" responses.

[9]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.

responses, we indicated customer income in our prompts (chosen to be the median income from the US Census, similar to the Fong et al. (2023) sample). For simplicity, while we were inspired by the Fong et al. (2023) design, we varied only brand, price, and fluoride content, and kept "whitening" constant in all prompts. We also only had GPT consider choice sets with two different brands.

For the remainder of our baseline GPT studies, we removed customer income from the baseline GPT prompts to make them more general,[10] and we also ran a full factorial conjoint on GPT, including the brand attribute. With respect to number of queries for each consideration set, while we initially collected 300 responses for each consideration set in our pilot studies, we collected only 150 responses for our main studies and then only 50 responses, as we learned that GPT's responses are consistent at smaller sample sizes (see Online Appendix G for details).

2.3.3 Exploring Heterogeneity in Preferences

Market researchers interested in customer heterogeneity often analyze survey responses from different subgroups based on demographics, behavior, or other characteristics. To emulate this type of analysis, we collect a set of demographic characteristics at the end of our human conjoint surveys. We use this data to construct different customer segments and compare WTP between these groups. We then use GPT to simulate responses from equivalent subsamples of customers by adapting the prompt to include relevant demographic features (e.g., income, gender, or race). We compare the preferences implied by responses of a particular human segment to its corresponding GPT segment.

We collect self-identified information on gender, age, race/ethnicity, political affiliation, household income, and education level in each of our human studies. Our participant pool is similar to the US population in terms of gender and race, but tends to be younger, more educated, more likely to identify as Democrats, and lower income compared to the general population. See Table D.1 in the Online Appendix for details.

2.3.4 Fine-tuning with Relevant Context

Practitioners conducting market research on new products, features, or customer segments may have access to previous studies on similar or adjacent products, features, and customer segments. We incorporate these prior surveys by fine-tuning GPT and in doing so we modify GPT's weights directly.[11] Specifically, our approach to fine-tuning is to transform the responses

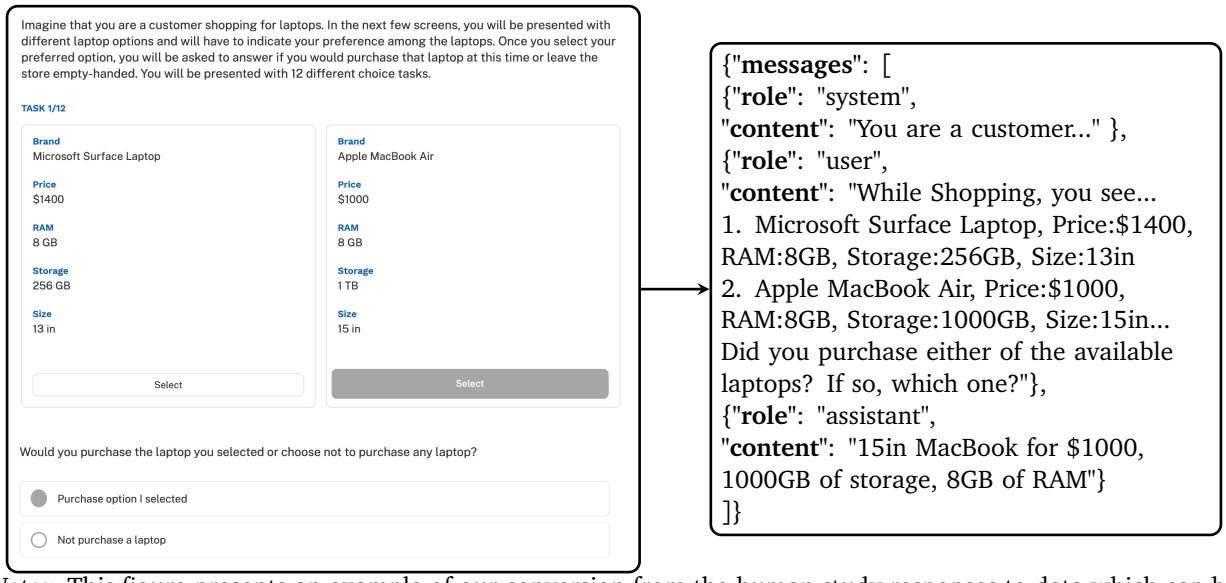
[10] We revisit this approach when exploring heterogeneity, as described in Section 2.3.3.

[11] In theory, there are many ways to provide contextual information from prior market research to GPT. One could, for example, directly include quoted customer responses, or summaries of survey data, in preambles ap-

from the prior human surveys into a new set of prompts and response pairs that are identical in format to our GPT prompts, which we describe in Section 2.3.2. Unlike our typical prompts, where we simply query GPT and collect the response, in our training data we provide pairs of prompts *and* responses, constructed based on the choice tasks and the responses of the participants in the study. We show an example of this construction process in Figure 1. Following this format, we use our entire existing human study to fine-tune GPT. For example, for a study with 300 participants and 12 questions each, we use 3,600 responses to fine-tune GPT.

In our fine-tuning studies, we use human responses from Study 1A (conventional toothpaste) to fine-tune GPT and evaluate its responses to Study 1B (toothpaste with new and unusual flavors), and similarly, we use responses from Study 2A (laptop) to fine-tune GPT and evaluate its responses to Study 2B (laptop with a new feature) and 2C (tablet with different levels of the same attributes as the laptop).

Figure 1: Converting Human Survey Response to Fine-tuning Input Material



Notes: This figure presents an example of our conversion from the human study responses to data which can be provided to GPT for fine-tuning. The human respondent in this case chose to purchase the Macbook Air, which we convert to a text response in the assistant's "content" field to send to GPT. Our fine-tuning data, on the right hand side, includes our description of the setting (a customer who is randomly selected while shopping), the available products, and the customer's choice (complete system and user prompts appear in Online Appendix E).

pended to the beginning of GPT prompts. Alternatively, as has become popular in agent-based models, one may provide survey materials as documents that GPT has access to and use retrieval-augmented generation (RAG) (Lewis et al., 2020), with the aim of producing responses informed by the additional information.

3 Results

In this section, we present results from studies designed to assess the usefulness of GPT-based conjoint studies as a tool for market research. We begin with results that show how estimates of WTP derived from GPT responses compare with those derived from human surveys. Next, we present results on GPT's performance in representing consumer heterogeneity. Finally, we show that fine-tuning GPT with human surveys can result in GPT responses more consistent with those of humans', including on choices involving new product features.

3.1 Recovering Baseline GPT Preferences via Conjoint

Our first set of results evaluates the ability of GPT to produce responses that reflect consumers' WTP for product attributes without any supplemental contextual information beyond what is contained in brief survey prompts. Our goal here is to assess GPT as a tool for market researchers to generate a baseline estimate of the rank and magnitude of customers' WTP in settings where no historical surveys are available. For this, we present the results of three different exercises. First, we report GPT's implied preferences for basic attributes of toothpaste and deodorant, comparing GPT's output to results from humans in Fong et al. (2023). Second, we compare GPT's WTP for toothpaste, laptop, and tablet attributes to our human studies, focusing on attributes that existed prior to GPT's launch. Finally, we turn our attention to new features, comparing GPT's and humans' preferences for a built-in laptop projector and to new and unusual toothpaste flavors.

3.1.1 Pilot Studies

Our first set of results, which survey GPT's preferences for toothpaste and deodorant, are shown in Table 3. In these conjoint studies, we created the full set of options for each brand: three price levels for each of the attribute options, yielding a total of 36 configurations. As detailed in Table 2, our conjoint design is inspired by the choices in Fong et al. (2023). We use the top two brands for each product category, have three price levels for each of the goods, and two attribute levels for fluoride and aluminum content (with, without).

Here we present estimated utility parameters from our two studies, which were conducted in March 2023 using the text-davinci-003 model through OpenAI's GPT API, as well as the implied WTP for aluminum in deodorant and fluoride in toothpaste, which we estimate using a multinomial logit choice model. We note that LLM responses are simulation-based responses, and are therefore akin to counterfactual simulation responses. However, unlike most settings, we know neither the data used to estimate the LLMs nor the model structure, meaning we have

Table 3: Pilot Studies Conjoint Results

	Toothpaste		*Deodorant*	
	(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

Notes: Significance level (calculated as if data is from a random sample of consumers): 10% (*); 5% (**); 1% (***). Attribute="fluoride" in Columns 1 and 2 , and "aluminum" in Columns 3 and 4; Brand dummies are "Colgate" (1) and "Crest" (2) for toothpaste, and "Dove" (1) and "Speed Stick" (2) for deodorant.

no way to precisely characterize sampling uncertainty arising from the LLM. Although LLMs represent unique challenges for characterizing sampling variance, we calculate standard errors in Table 3 as if our data were generated by randomly sampled consumers. We view this as a useful baseline to give a general sense for the variation in the simulated responses, though we note that future work may wish to explore alternative approaches to inference in these settings. When presenting implied WTP values we choose to omit standard errors.

We estimate that consumers are willing to pay, on average, $3.40 to add fluoride to toothpaste and $0.91 to remove aluminum from deodorant.[12] These studies, which were modeled after those in Fong et al. (2023), compare favorably to that baseline. Fong et al. (2023) found that consumers are willing to pay $3.27 on average for fluoride and -$1.97 for aluminum, meaning that in both cases our estimates come within 50% of theirs. This is notable because, in addition to allowing us to match our estimates from GPT to a human sample outside of the surveys we ran ourselves, Fong et al. (2023) also showed that the estimates and market shares from their survey matched real market outcomes.

3.1.2 Main Studies

Now we move to our second set of results, which compare GPT to our new set of human studies. For these main studies we use GPT-3.5 Turbo. Broadly, we find that GPT performs well at approximating human preferences in some contexts, and struggles in others. First, we examine the toothpaste studies (Studies 1A and 1B). Our estimates for WTP for each attribute are illustrated in Figure 2. Our figures include 95% confidence intervals (CIs) for WTP estimated

[12]To evaluate the stability of estimates over different models of GPT, we repeated this exact toothpaste study design on GPT-3.5 Turbo. We found slightly higher WTP, at $4.42. We then also examined the WTP when the income of the customer is not mentioned, and found that it increases to $4.95, suggesting that GPT's "generic" customer is wealthier.

from human studies, but not for LLMs.

Our results from GPT in Study 1A very nearly rank all of the attributes correctly. We find that the brand value of Crest, relative to Colgate, is small relative to all other attributes, although we did find that GPT's WTP differed in sign from humans (the CI for humans includes zero). Estimates of WTP from GPT for cinnamon and strawberry toothpaste flavors (relative to the mint baseline) are both correctly signed and of a similar magnitude to estimates from our surveys of humans, although GPT's WTP for strawberry flavor is more negative than humans'.

The main difference between humans and GPT in this study is that GPT has higher WTP for fluoride. Our findings in Study 1B are similar, with GPT reflecting higher value for fluoride and greater distaste for non-mint flavors. At the same time, we note that the estimates of WTP for fluoride from the human study are comparable to the estimate from Fong et al. (2023) (estimates of humans' WTP for fluoride are $2.60 and $2.90 in Studies 1A and 1B, respectively). We also find that GPT's WTP for fluoride ($8.20 and $9.30 in Studies 1A and 1B) is almost twice as large as what we found when replicating the pilot study on GPT-3.5 Turbo (between $4.42 and $4.95). We believe that this is likely due to adapting the prompts to include the full factorial design (including brands), and to adding additional attributes to the prompt (flavors), which alter GPT preferences but not human preferences. We discuss the sensitivity of our results to the specific LLM used below (Sections 3.1.3 and 3.3).

Next, when evaluating new flavors in Study 1B, GPT differs from human data in assessing the sign of WTP for our brand new toothpaste flavors, cucumber and pancake. This is somewhat surprising, given GPT's consistent assessment of WTP for strawberry and cinnamon flavors, which exist but are significantly less popular flavors for toothpaste than mint. Based on GPT's responses, we believe some of the differences can be explained by its preference for new product trials, as we found that GPT was more likely to opt out of purchasing (relative to Study 1A) when it was not offered the new toothpaste flavors (see Online Appendix F for selected prompts illustrating this point). We discuss these results in more depth below.

Next, we move to our results on consumer electronics, which are presented in Figure 3. In Study 2A, we see that GPT is consistent with human data in ranking customers' WTP for storage and RAM, and produces estimates within 64%–90% of our human sample's estimated WTP. The same is true in Study 2B, in which we survey a new sample of human respondents. In contrast, in both studies, GPT's responses imply a negative WTP for the Macbook brand relative to the alternative, whereas humans in both studies suggest a significantly higher preference for the brand. We see a similar comparison in Study 2C (Tablets), where GPT also has a relatively low preference for iPads. In this case, GPT does however sign this value similarly to human

data and is within $50 (67%) of our human estimate. However, GPT's estimates of WTP for other attributes (Storage, RAM, and Size) are substantially larger than humans' WTP for these attributes (70%–312% larger, and outside of the 95% CI). When exploring the new feature of a "built-in projector" in Study 2B, we again find that GPT's WTP for the new feature appears to be higher than humans' WTP for this feature by almost 3X.

With respect to screen size in laptops, we find that, in contrast to humans who prefer a larger screen size, GPT suggests a negative WTP in both Study 2A and 2B. Upon further exploration, we found that this is due to our design choices – it turns out that GPT prefers a laptop screen size of 13 inches (over both smaller and larger sizes), and because our options only included 13 and 15 inches it generated a negative WTP for this attribute.[13]

Overall, we find that for many of the attributes, GPT-3.5 Turbo produces WTP estimates that are of similar order of magnitude and sign as our human sample; however, there exist some substantial differences for some attributes. Therefore, at their baseline form, GPT's responses can be seen as a useful signal, but not as a replacement to human responses, as they may lead to wrong conclusions especially about new features (e.g., GPT results suggest that humans will like the "pancake" toothpaste flavor). In Section 3.3 we examine whether providing additional data may help GPT produce responses more similar to those of humans.

3.1.3 Robustness to the Choice of LLM

In this section we examine whether and to what extent the reliability of LLMs for market research depends on the choice of LLM, which we view as closely related to the stability of results from a single LLM over time. We address this question by replicating our analysis of our Study 2A[14] using GPT-4o, Claude 3 Haiku, Claude 3.5 Sonnet, and LLaMA-3 (8B and 70B).[15] Comparisons between these models (Figure 3a) highlight some notable differences. In general, it appears that responses from GPT-3.5 Turbo and GPT-4o much more closely approximate the responses we see in our human sample than do those of Claude and LLaMA-8B. In general, LLaMA-8B understates human preferences significantly (perhaps because it is the smallest model); Claude suffers less from this problem, but its preference for Macbooks appears to far exceed humans' and other models' preferences, and its preference for screen size is the far-

[13] To examine this, we ran a test that kept most attributes stable and only varied prices, brands, and sizes, and found that the ideal size for GPT was 13 inches. We later also tested this for GPT-4o and found that it prefers a 15-inch screen (when screen sizes varied between 11 and 17 inches).

[14] We evaluate robustness using our laptop studies because they include a larger set of attributes than our toothpaste studies.

[15] In our replication studies, we use 50 repetitions per prompt (total of 288,900) for each LLM, except for the LLaMA 70B model that only has 2 repetitions per prompt (11,556 prompts total) due to the computational intensity of running this Open Source model locally.

thest from humans among all of the models'. Even between GPT-4o and GPT-3.5 Turbo there are some notable differences, such as their assessments of WTP for the Macbook brand, and their preference for screen size.

In practice, researchers likely want any LLM augmentation of market research to be grounded in existing data whenever possible, and only supplemented by the LLM when extrapolation beyond the available data is needed. Given sufficient contextual data in the form of fine-tuning, our hypothesis is that some of the differences between LLMs would shrink. We discuss the impact of such fine-tuning approaches on the insights gleaned from GPT surveys, both in-context and out-of-context, in Section 3.3.

3.2 Examining Heterogeneity in WTP

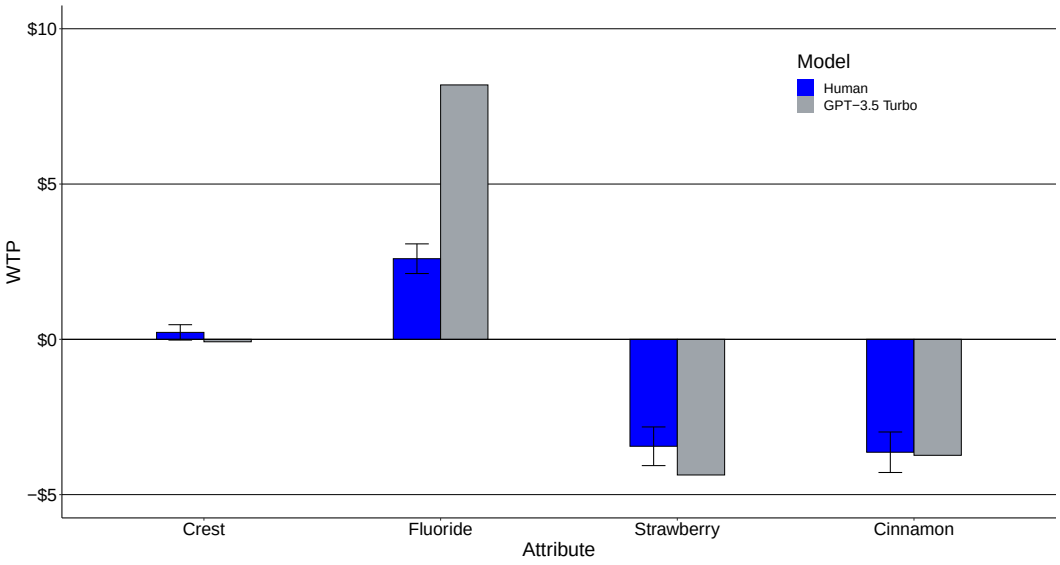
A critical element in survey design is defining the target population. We are interested in understanding GPT's ability to reflect the preferences of different populations. To this end, we rely on the population of our human studies who were Prolific participants residing in the US and aged 18 or older. To better understand GPT's ability to mimic the preferences of different populations, we collected demographic data from humans (see Online Appendix D for details), estimated WTP separately for each demographic group, and then submitted queries to GPT corresponding to each group. In particular, we modified our prompts to indicate customer income, gender, or race. See Online Appendix E for details.

We use Study 2A for our investigation of customer heterogeneity. In our human studies, we found that within a particular demographic split (e.g., income), there was relatively little variation in WTP for RAM or storage or the fraction of those choosing not to purchase (which was around 50% across all groups, aside from Black participants, whose opt-out rates were only 37%). Meanwhile, GPT reflected large variation between demographic groups for these measures. For example, for the income groups below $50,000 and above $150,000, WTP for RAM among humans was $17.80 and $19.40, respectively, whereas the GPT-based estimates were -$4.40 and $99.60, respectively. GPT also had an opt-out rate of 95% for the lower income group and 34% for the higher income group, whereas the corresponding opt-out rates among humans were 49% and 47%. A comparison of human- and GPT-implied WTP across groups is available in Online Appendix G.[16]

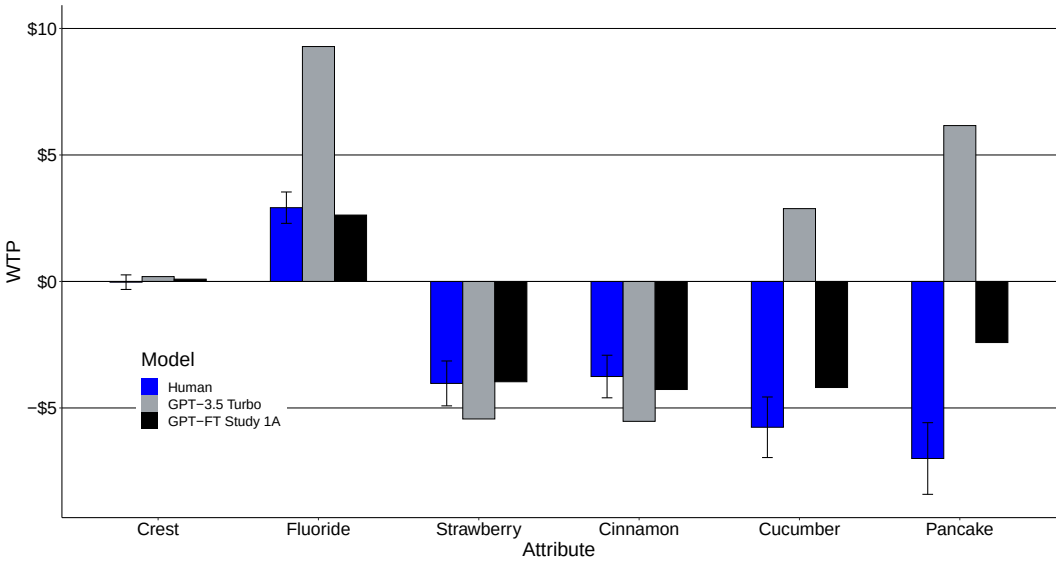
In our human surveys we find that WTP for screen size and for the Macbook product (relative to Surface), varied more substantially between groups relative to other attributes. For clarity

[16]Throughout our analyses, GPT always implies a negative price coefficient, except for the case of higher income (above $150,000) customers, where the price coefficient is positive.

Figure 2: Results of Toothpaste Studies



(a) Study 1A: Toothpaste

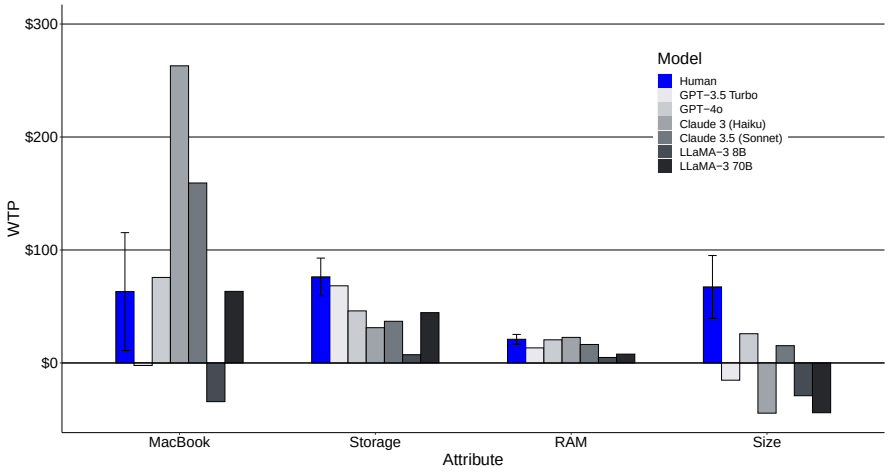


(b) Study 1B: Toothpaste with New Flavors

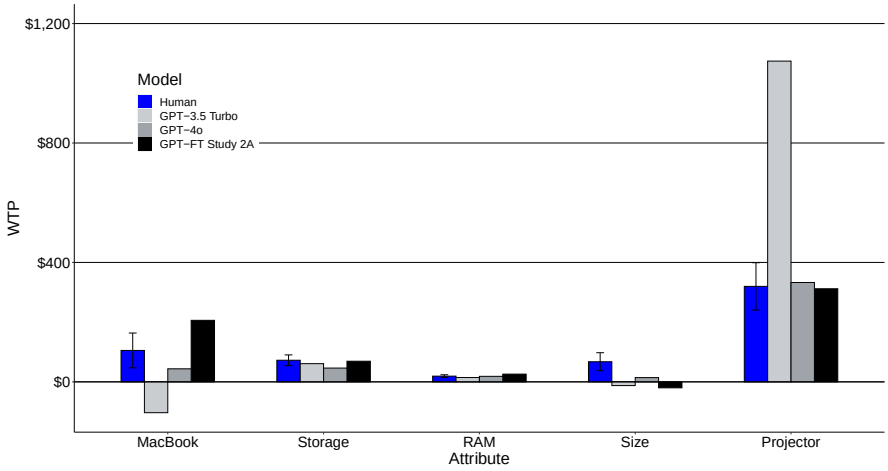
Notes: Estimated WTP for humans and LLMs for Studies 1A and 1B. Brackets indicate 95% confidence intervals on our estimates from the human study. GPT-FT represents a fine-tuned GPT model.

of presentation, we focus our figures on these two variables. Figure 4 presents the comparison between the estimated WTP based on humans and on GPT for the different groups. An interesting patterns emerges for comparisons between groups by income, gender, politics, and race and ethnicity (but not for education or age) for Macbook WTP (but not for screen size): while GPT estimates are of different magnitudes and signs compared to those from humans, GPT

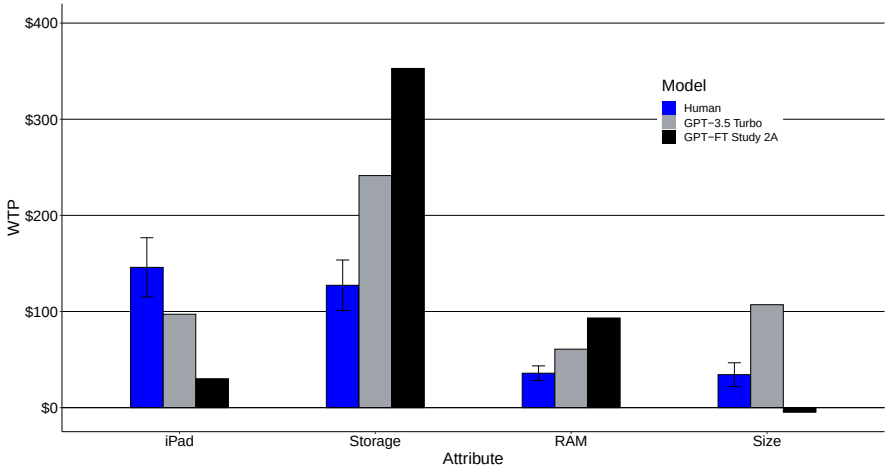
Figure 3: Results of Laptop and Tablet Studies



(a) Study 2A: Laptops



(b) Study 2B: Laptops with Built-In Projector



(c) Study 2C: Tablets

Notes: Estimated WTP for humans and multiple LLMs for Studies 2A, 2B, and 2C. Brackets indicate 95% confidence intervals on our estimates from the human study. The reported WTP for screen size and RAM are for one unit (inch/GB), and for Storage it is for 128GB. GPT-FT represents a fine-tuned GPT model.

does seem to reflect consistent ordering between the groups. For example, in our human sample WTP for Macbook increases from $12.50 to $225.60 with the increase in income bracket, and GPT's WTP increase from -$2.60 to $722.52. However, it seems that GPT generalizes this order to all the attributes (i.e., to storage and RAM), differently from what we see in data from humans. Focusing on gender, which has the largest split in our humans samples (44% of respondents identified as male and 54% as female) and smallest CI, GPT WTP estimates for Macbook are within the CI, but this is not the case for screen size or other attributes.

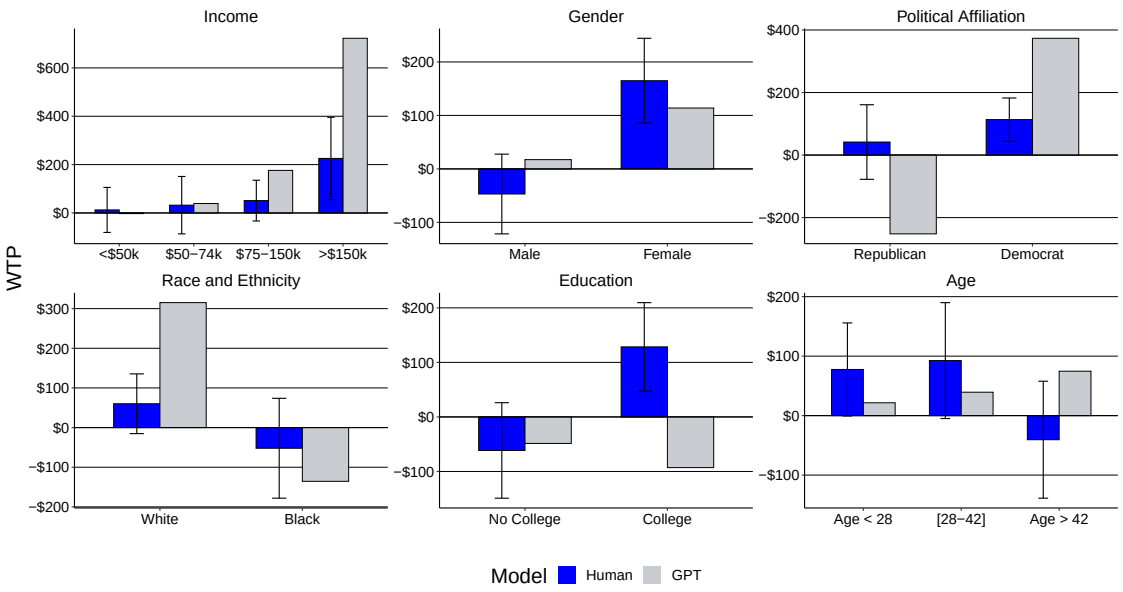
Overall, we find that GPT reflects some differences between customer groups found in human samples but fails to reflect others. Specifically, it does not reflect preferences of any particular group very well. This is in contrast to what we find in Section 3.1, where "off-the-shelf" GPT-3.5 Turbo provides similar average WTP estimates across the entire population for several product attributes. In the next section, we investigate whether supplementing GPT with additional information may improve its ability to reflect preferences derived from human data.

3.3 Supplementing GPT-3.5 Turbo with Fine-Tuning

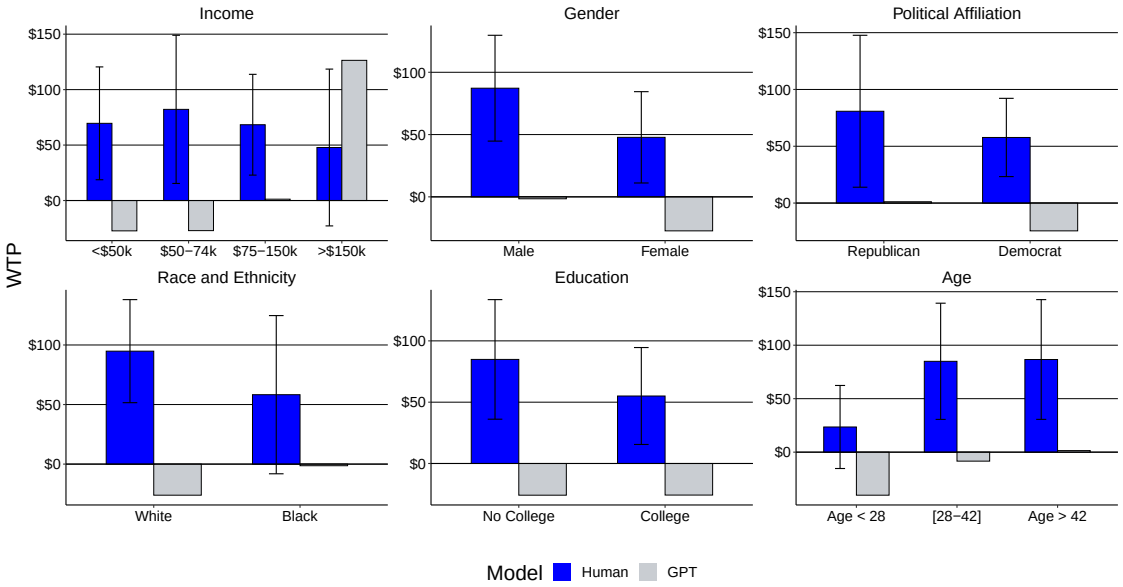
Having documented encouraging but mixed results from surveying GPT without providing any additional information, we now move to exploring how GPT responds to being fine-tuned with related surveys from a target population. We focus on two questions as our measures of success: (1) Does fine-tuning yield better alignment between GPT and humans on the products and features included in the fine-tuning data? (2) How does GPT perform on out-of-sample questions, including both similar but related product categories and new product features?

To study both these questions, we fine-tune GPT with human survey data from one study, and then query GPT with questions from another study. Beginning with the first of our two questions, we fine-tune GPT with a study on existing products (toothpaste or laptops) and then query it about products with new features (new flavors or a built-in projector). We look first to our results from Study 1B, in which we fine-tuned GPT using Study 1A (see Figure 2b), and Study 2B, where we fine-tuned GPT-3.5 using our data from Study 2A (see Figure 3b). In both cases we find improvements on some of the attributes that were included in our fine-tuning data. Relative to baseline GPT, WTP for fluoride decreases to a similar level as humans after fine-tuning. Notably, fine-tuning GPT using Study 1A increases the accuracy of GPT's WTP for cinnamon and strawberry flavors relative to both of our study samples in Studies 1A and 1B. This finding is particularly encouraging, as it suggests that providing GPT results from a survey on one sample of humans helped to predict a second sample's responses to similar questions. Similarly, the sign of preferences for Macbooks flips in Study 2B after fine-tuning, and our

Figure 4: Study 2A WTP by Demographic Groups



(a) Macbook Preference (over Surface)



(b) Screen Size

Notes: Estimated WTP for Macbook relative to Surface and for screen size, estimated separately for six demographic splits. Brackets indicate the 95% confidence interval on the estimated WTP from our human sample.

estimated WTP for storage improves as well.

On the other hand, we find some cases where fine-tuning had less impact on GPT's WTP. The best example of this is laptop size, which baseline GPT-3.5 (as well as other LLMs), unlike

humans, identify as a negative. Fine-tuning GPT with human responses from Study 2A does not improve alignment. One possible explanation for this issue is that GPT's training data introduces a strong signal that a 15-inch screen size is a negative feature in laptops.

To address our second question, we study GPT's ability to extrapolate from data we use to fine-tune it. First, we study extrapolation to a separate but related product category. In Study 2C, we fine-tune GPT with survey data about laptops but query it about tablets (see Figure 3c). We find that fine-tuning *increases* the distance between human- and GPT-generated WTP estimates for brand, storage, and RAM. Fine-tuning reduces the absolute difference in WTP for size, but flips the sign of this preference relative to humans. Yielding the incorrect sign of WTP may well be more detrimental than overestimating WTP. Overall, we do not find supportive evidence of GPT's ability to extrapolate to other product categories after fine-tuning (even when they have the same attributes). We propose two reasons why GPT is not better aligned with humans in this case: i) Attribute values for tablets are the same or lower than the worst levels for laptops (e.g., 16 GB of RAM is the largest available for tablets but the lowest for laptops), which may impact GPT's ability to learn about these levels; ii) even before fine-tuning, GPT's estimates for tablets were the most divergent from human samples.

We find more encouraging results when we turn our attention to extrapolation *within category*. We withheld information on three attributes from GPT during fine-tuning: cucumber and pancake flavors for toothpaste, and a built-in projector for laptops. Turning to Study 1B first, fine-tuning GPT with data from Study 1A flips the signs of WTP for both of our new flavors from positive to negative, matching the sign from our human sample. Our post-fine-tuning estimates of WTP for the cucumber flavor are within 30% of our estimates from humans, although they are still outside of the corresponding CI. Our WTP estimates for cinnamon-flavored toothpaste are within the CI for the estimate from humans in Study 1A, and fine-tuning Study 1B with Study 1A also brings GPT's WTP for cinnamon within the CI in 1B. Our WTP estimates for pancake-flavored toothpaste (relative to mint) are less similar: we derive a WTP of -$2.40, whereas our human sample indicates a WTP of approximately -$7.00. Still, in this case a fine-tuned GPT would help a market researcher correctly assess that both new flavors are unlikely to be popular with the target population, owing to the dramatically reduced distance between GPT's and humans' WTP estimates.

We see another example of successful extrapolation within category when we look at Study 2B, our survey concerning laptops with a new projector feature, which we fine-tuned with our data from Study 2A (laptops without projectors). In the baseline version of this study, GPT overestimated customers' WTP for a projector by more than a factor of three. After fine-tuning,

the two estimates align very closely. In this study, then, fine-tuning was essential to developing reasonable estimates, and it produced not only the correct sign but also the correct magnitude of WTP for the target audience. We close by noting that we replicated our study on GPT-4o, without fine-tuning, and this exercise produced a WTP very similar to our humans and to our fine-tuned GPT. Although this is only one example, this supports our hypothesis that fine-tuning may help to reduce some of the heterogeneity in estimates across LLMs.

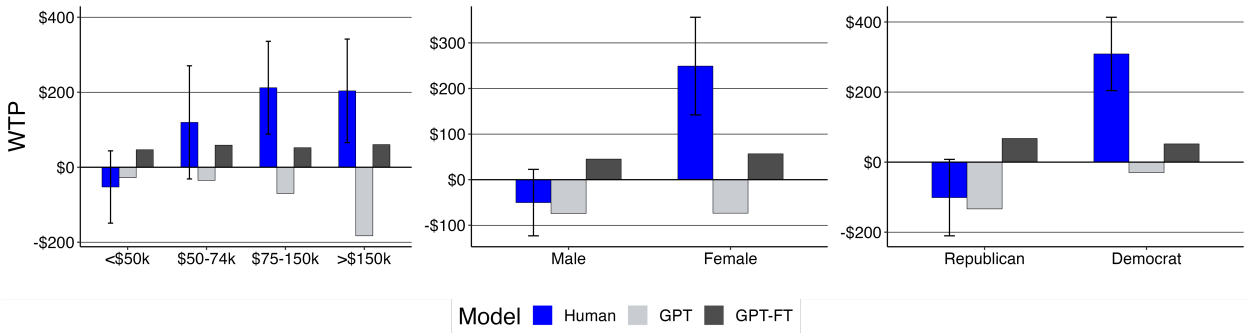
Finally, we also examine whether fine-tuning may be able to improve our customer heterogeneity results. To do this, we again fine-tune GPT with Study 2A responses, but we modified the prompt to include the income bracket and gender of the customer based on our survey responses. We then query this fine-tuned GPT with our questions from Study 2B where we separately ask about customers from different income levels, genders, and political affiliations. Overall, we find that fine-tuning GPT with demographic information generated WTP estimates that are more similar to the average estimates from the human studies, and also correctly changed the sign of preferences for Macbooks (relative to customers with $50-75k in income) across the income distribution.

However, GPT was unable to meaningfully recover the differences across the different demographic groups, both for attributes that existed in Study 2A and for the new attribute, even when those differences were statistically significant (e.g., lowest and highest income brackets, gender, and political differences in Figure 5a). For example, the mean WTP for Macbook among human males was -$50 and for females it was $249, but the fine-tuned model suggests WTPs of $45 and $57, respectively. For the projector, the fine-tuned model suggests a WTP of between $170 and $196, which is outside of the CI for the majority of the human sub-samples. Figure 5 presents the results for WTP for Macbook and projector, and Online Appendix G.2 includes additional details about the procedure and presents the WTP estimates for all the different attributes. Given these results, a marketer interested in using our method to study the WTP of particular customer segments may have to train a separate model for each customer segment they are interested in (using sufficient past surveys on that population).

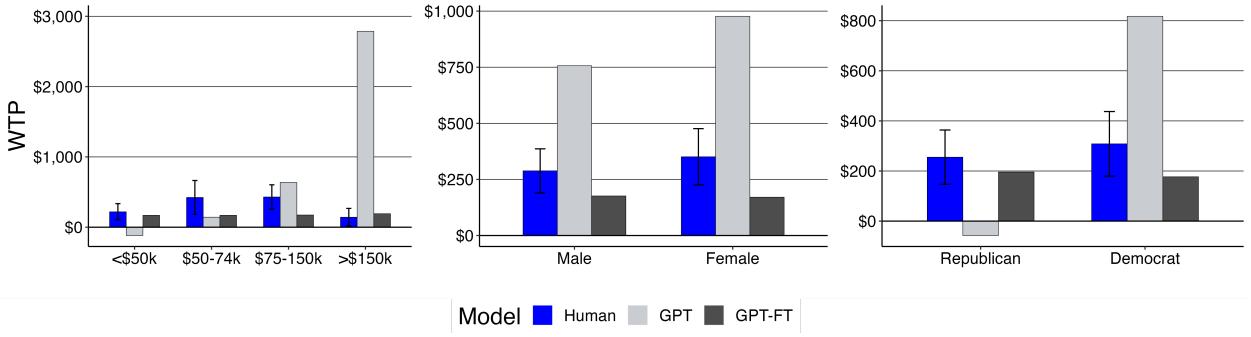
3.4 Potential for Scalability of Our Fine-Tuning Approach

The results above leave open the question of how this approach would scale in practice. We have only fine-tuned GPT with a single study at a time, whereas market research firms could have dozens of studies relevant to a class of products. We intuit that certain relationships in a dataset should be fundamentally easier for an LLM, or any machine learning model, to “learn" than others. In small samples especially, LLMs may more easily learn simple moments such

Figure 5: Study 2B WTP by Demographic Groups



(a) Macbook Preference (over Surface)



(b) Built-In Projector

Notes: Estimated WTP heterogeneity for Macbooks, relative to Surface, and for projectors in laptops, estimated separately for gender, income, and political affiliation. Brackets indicate the 95% confidence interval on the estimated WTP from our human sample. GPT-FT represents a fine-tuned GPT model.

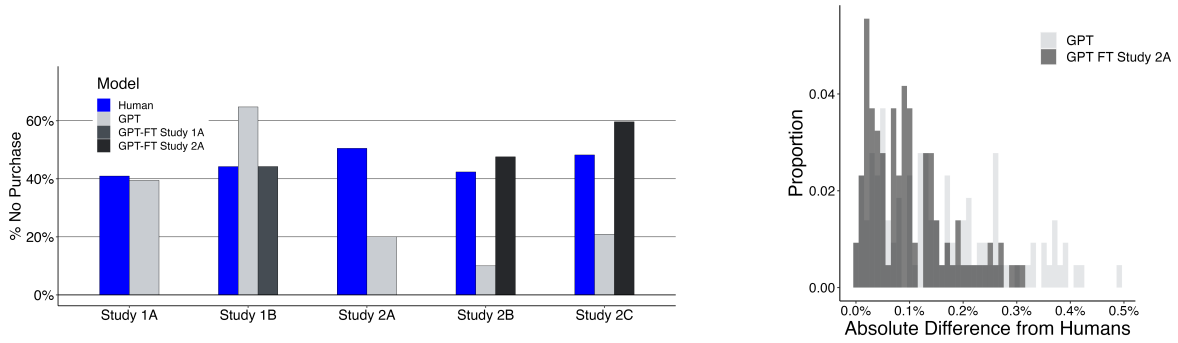
as market shares rather than estimate WTP, which is a complex and highly model-dependent function of the data.

Fine-tuning GPT on our target customer base merely modifies the model's weights to better fit the data we have provided. GPT then uses its other implicit knowledge to respond to our survey prompts and extrapolate from our data. Intuitively speaking, if we can show that our single studies are enough to more closely match certain moments between GPT- and human-generated responses, we should expect that, consistent with typical scaling rules for model training, larger amounts of fine-tuning data should help GPT learn more complex relationships in the data.

In Figure 6 we plot two sets of model-free outcomes to demonstrate intuitively that GPT learns these relationships in the data. In Figure 6a, we plot outside option shares from humans and GPT in each of our studies, including before and after fine-tuning where relevant. Here we can see that in each of the three studies in which we fine-tuned GPT, fine-tuning makes GPT choose

the no-purchase option at rates more like our human participants. This is true even in Studies 1B and 2B, where the choice set in the new study differs from the fine-tuning data (because they include new features), indicating some level of successful extrapolation. In Figure 6b, we restrict our GPT-generated data from Study 2B to only the exact set of questions which were used to fine-tune GPT (i.e., those which were included in the Sawtooth human survey version of Study 2A), and compare the similarity of choice shares between GPT and humans before and after fine-tuning. Again, we see that fine-tuning generated intuitive improvements in accuracy on these empirical moments relative to our human benchmark. These comparisons suggest that our approach may be expanded by research firms over time.

Figure 6: Results from Fine-tuning



(a) Outside Option Shares

(b) Study 2B Market Shares

Notes: This figure presents selected moments of our data in which we can see fine-tuning improve GPT's ability to replicate human choices. Figure (a) presents the outside (no purchase) option for each study, comparing GPT before and after fine-tuning to the closest human comparison group. Figure (b) presents the absolute difference in choice shares between GPT and humans (in percentage points), among the specific tasks which were provided to GPT via fine-tuning. GPT-FT represents a fine-tuned GPT model.

4 Conclusion

This paper studies the usefulness of GPT and other LLMs for market researchers. We focus on the practical problem of a market researcher who would like to learn something about customer preferences in a new setting but may not have the time or budget to run new studies with human subjects. With this setting in mind, we designed a custom set of studies that allow us to mimic the market researcher's existing data resources and to control the amount and type of information GPT has about the context under study. We then compare GPT's responses to survey questions with those of humans'.

We find that in many cases GPT responds to market research questions similarly to humans even without any fine-tuning. In a pilot study, we found that GPT's implied WTP for certain

consumer goods was close to estimates of WTP from a recently published study, and the same was true for several product attributes we analyzed in our new custom studies. When we supplemented GPT via fine-tuning, we found that GPT became better aligned with human responses even for new-to-world product features. However, it struggled to extrapolate to product categories outside of the fine-tuning data, and fine-tuning in this context worsened GPT's performance. We also saw mixed results when we studied customer heterogeneity, where fine-tuning often improved GPT's alignment with human responses on average but was unable to generate comparable heterogeneity. We envision marketers using our approach to help them test out and narrow down new feature ideas before testing them with humans, as opposed to replacing their study subjects with LLMs.

Our work surfaces some limitations of using LLMs for market research. 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 customer heterogeneity. 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 (although our methodology can mitigate this issue by augmenting the LLM with human studies). Additionally, our work emphasizes the sensitivity of GPT to how prompts are worded (see Online Appendix A for examples we came across, although our results were overall robust to prompt phrasing). Furthermore, rapid development cycles and frequent introduction of new LLMs necessitates evaluating baseline responses of each LLM release. We already see in various examples (see Figure 3) that different LLM provide similar estimates, and we believe that our fine-tuning approach can further improve the reliability of each LLM.

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-4o) 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.

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.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 1–16.

Burnap, A., Hauser, J. R., & Timoshenko, A. (2023). Product aesthetic design: A machine learning augmentation. *Marketing Science*, *42*(6), 1029–1056.

Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2023). Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951*.

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*.

Goli, A., & Singh, A. (2024). Frontiers: Can large language models capture human preferences? *Marketing Science*.

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.

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.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Li, P., Castelo, N., Katona, Z., & Sarvary, M. (2024). Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science*, *43*(2), 254–266.

Moshary, S., Shapiro, B., & Drango, S. (2023). Preferences for firearms. *Available at SSRN*.

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.

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.

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

Appendix

A Code Example

The following Python code was used to collect GPT responses for Study 2A:

```python
# Import packages (install first if needed)
from openai import OpenAI
import itertools
import pandas as pd
import time
import os
from dotenv import load_dotenv
from concurrent.futures import ThreadPoolExecutor, as_completed
load_dotenv()

# Include API Key provided by OpenAI (www.platform.openai.com)
api_key = os.getenv('api_key')
client = OpenAI(
    api_key = api_key
)

# Functions for interacting with OpenAI API (prompting, extracting responses, etc.)
def request_with_retry(model, prompt, max_retries=5, base_delay=1):
    for attempt in range(max_retries):
        try:
            response = client.chat.completions.create(
                model = model,
                messages = [
                    {
                        "role": "system",
                        "content": "You are a customer. You are selected at random while shopping for laptops to
                            participate in a survey. The interviewer will describe the options you saw while
                            shopping and ask you to report which option you chose to purchase. Whenever two
                            options are shown, you can also choose a third option which is not to purchase
                            anything that day."
                    },
                    {
                        "role": "user",
                        "content": prompt
                    }
                ],
                temperature = 1,
                max_tokens = 256,
                top_p = 1,
                frequency_penalty = 0,
                presence_penalty = 0
            )
            return response
        except Exception as e:
            if attempt == max_retries - 1:
                raise
            else:
                sleep_duration = base_delay * (2 ** attempt)
                print(f"Error: {e}. Retrying in {sleep_duration} seconds.")
                time.sleep(sleep_duration)

def generate_prompt(product_1, product_2):
    prompts = [
        f"{product_1[0]}, Price: ${product_1[2]}, RAM: {product_1[3]}GB, Storage: {product_1[4]}GB, Size: {product
            _1[1]}in",
        f"{product_2[0]}, Price: ${product_2[2]}, RAM: {product_2[3]}GB, Storage: {product_2[4]}GB, Size: {product
            _2[1]}in"
    ]
    return f"While shopping, you see two options:\n \n1. {prompts[0]}\n \n2. {prompts[1]}\n \nYou also have the
         option not to purchase a laptop.\n \nDid you purchase either of the available laptops? If so, which one?"

def process_set(model, index, set, iterations):
    responses = []
    product_1 = set[0]
    product_2 = set[1]
    prompt = generate_prompt(product_1, product_2)
    response = request_with_retry(model, prompt)
    answers = {
        'set': index + 1,
        'first_product': product_1,
        'second_product': product_2,
        'prompt': prompt,
        'response': response.choices[0].message.content.strip()
    }
    responses.append(answers)
    return responses

def query_gpt(model, iterations, brands, size, prices, ram, storage):
    products = list(itertools.product(brands, size, prices, ram, storage))
    sets = [list((p1, p2)) for p1 in products for p2 in products if p1 != p2]
    responses = []
    for index, set in enumerate(sets):
        response_data = process_set(model, index, set, iterations)
        responses.extend(response_data)
```

```
    sorted_responses = sorted(responses, key = lambda x: x['set'])
    return sorted_responses

# Running Study 2A with GPT 3.5-Turbo
if __name__ == '__main__':
    model = 'gpt-3.5-turbo-0125'
    iterations = 150
    brands = ['Microsoft Surface Laptop', 'Apple MacBook Air']
    prices = [1000, 1200, 1400]
    size = [13, 15]
    ram = [8, 16, 32]
    storage = [256, 512, 1000]

    output = query_gpt(
        model = model,
        iterations = iterations,
        brands = brands,
        prices = prices,
        size = size,
        ram = ram,
        storage = storage
    )
    df = pd.DataFrame(output)
    df.to_csv('output/Study1_output.csv', index=False)
```

B Prompt Example

A prompt that was used to create GPT responses for one consideration set in Study 2A:

> "**System message:**
>
> You are a customer. You are selected at random while shopping for laptops to participate in a survey. The interviewer will describe the options you saw while shopping and ask you to report which option you chose to purchase. Whenever two options are shown, you can also choose a third option which is not to purchase anything that day.
>
> **User message:**
>
> While shopping, you see two options:
>
> 1. Microsoft Surface Laptop, Price: $1400, RAM: 8GB, Storage: 256GB, Size: 13in
>
> 2. Apple MacBook Air, Price: $1000, RAM: 8GB, Storage: 1000GB, Size: 15in
>
> You also have the option not to purchase a laptop.
>
> Did you purchase either of the available laptops? If so, which one?"

Online Appendix

A 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 larger outside option shares. Importantly, conditional on making a choice, the implied marked shares are similar between the two types of phrases. These differences in prompting were especially crucial for conjoint studies. 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.

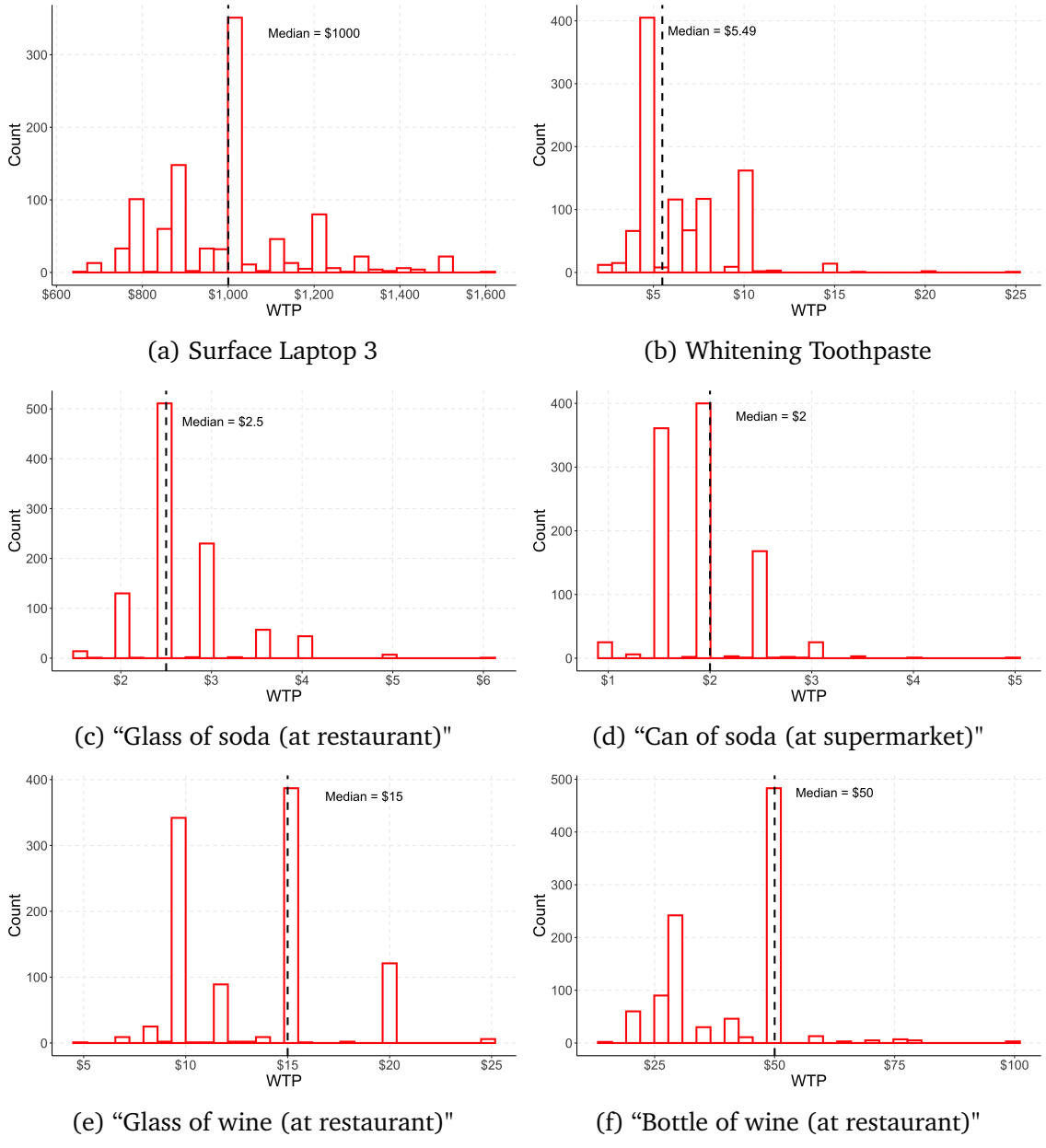
B Recovering Realistic WTP For Products

We explore whether asking GPT directly for 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). We follow our typical prompts, where a customer is randomly selected while shopping for a product and describe a good, but we omit the price of the good. We also adjust the prompting questions and ask: *"The custoemr is asked: What is the maximum price you would be willing to pay for [the good]? please give a single price as you answer."*[17] Figure B.1 reports our results. We begin by plotting the distribution of WTP for a Surface Laptop 3 with the specifications: Intel Core i5 processor, 8GB RAM, 13.5in screen, and 128GB Storage Drive. The median implied WTP for the Surface Laptop 3 is $1,000, similar to its market price.

Next, we use general descriptions to elicit WTP for a good, rather than WTP for a particular brand. In Figure B.1b we solicit WTP for a "whitening toothpaste." We then ask for the WTP for a glass of soda (Figure B.1c) and a glass of wine (Figure B.1e) at a restaurant, and find that the median WTP for wine is six times higher than for soda (median of $15.00 compared to $2.50). Finally, we also demonstrate that the WTP for these items depends on the size and the context, as the WTP for soda can at a supermarket is lower than at a restaurant (Figure B.1d) and the WTP for a bottle of wine is higher than for a glass of wine (Figure B.1f).

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

Figure B.1: 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)"

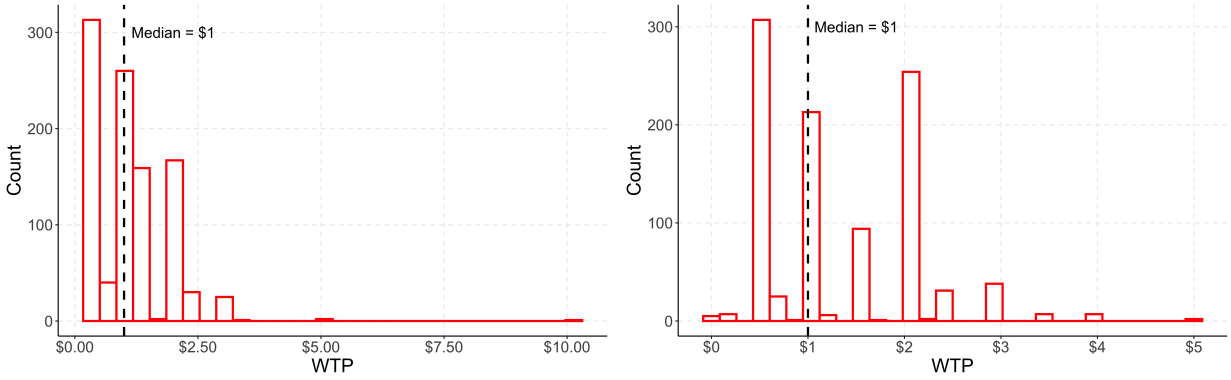
C Alternative Product Attributes WTP Solicitation Approaches

In addition to the conjoint approach used in the main text, we also tested two different strategies for eliciting WTP: a direct solicitation approach, and an alternative indirect elicitation approach. We note that direct solicitation of WTP for product attributes is not the typical method for human-based market research, as humans' ability to quantify these measures is limited. Instead, conjoint analysis is often used to derive WTP from survey responses. However, GPT may be more sophisticated than humans and may be able to calculate or infer WTP for attributes

from other training data. Therefore, we test the its ability to provide WTP for attributes directly or in a relatively simple indirect method before moving on to a conjoint study. We report the exact prompts we used for these studies in Online Appendix E.1.

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 (deodorant), we offer Colgate whitening toothpaste (Dove scented deodorant) and ask GPT how much more it would be willing to pay for the option with fluoride (without aluminum) over the option without fluoride (with aluminum).[18] Figures C.1a and C.1b report the results. The median and average WTP is \$1.00 and \$1.20 for fluoride and \$1.00 and \$1.30 for "no aluminum."

Figure C.1: Willingness to Pay for Attributes – Direct Solicitation



(a) Fluoride in toothpaste

(b) No aluminum in deodorant

Indirect elicitation The indirect elicitation approach consists of two steps. First, we use GPT to estimate the demand for the good with and without the attribute by asking GPT to make a choice between two toothpaste to generate two demand curves. Then, we compare these 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 (the focal good does not contain fluoride, but the reference good priced at \$4.00 does). Then, we compare the demand curve for the toothpaste with fluoride with the demand curve for the toothpaste without fluoride (see Figure C.2a 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 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 C.2a. For example, when

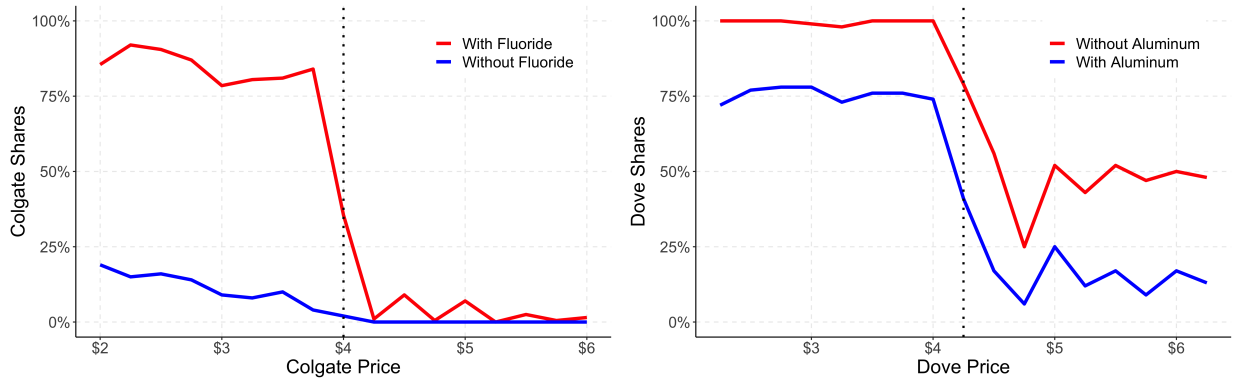
[18] In WTP queries, we query GPT 1,000 times for each product.

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 C.1a.

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 C.2b. As can be seen in the figure, GPT prefers aluminum-free deodorant. 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 C.2a, 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 C.2b 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 reported market shares are 34.7% for Crest versus 33% for Colgate) and a significant preference for Dove (the reported shares are 46% for Dove versus 22% for Speed Stick).

Figure C.2: Willingness to Pay for Attributes – Indirect Elicitation



(a) Toothpaste, Demand curve

(b) Deodorant, Demand curve

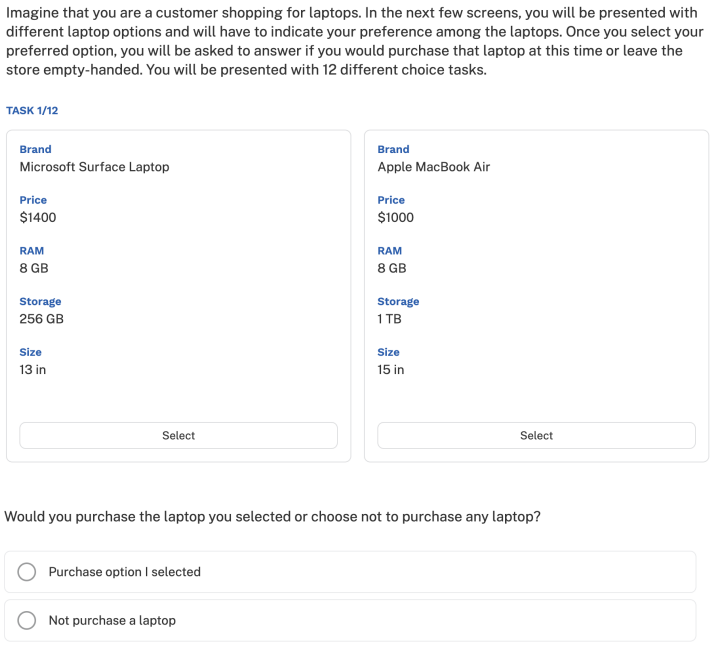
D Human Survey Materials

D.1 Surveys

All the human surveys were carried using the Sawtooth Discover software, which balances attributes across choices and selects the attributes to be nearly orthogonal all all configurations. In each of the studies, participants had 12 choice tasks where 2 product configurations were

presented. All of the studies had a similar preamble that said: *"Imagine that you are a customer shopping for [product]. In the next few screens, you will be presented with different [product] options and will have to indicate your preference among the [products]. Once you select your preferred option, you will be asked to answer if you would purchase that [product] at this time or leave the store empty-handed. You will be presented with 12 different choice tasks."* After each choice tasks, participants were asked to answer a question: *"Would you purchase the [product] you selected or choose not to purchase any [product]?"* Figure D.1 shows a sample choice question from Study 2A. While we allowed participants to first specify a choice, and only then decide if to purchase or not, our WTP analyses was based only on the choices conditional on purchase, similar to the GPT studies.

Figure D.1: Study 2A



For Study 1B and Study 2B, we added an additional sentence to describe the new features after the first sentence. In Study 1B, we added: *"Some toothpastes have new flavors: pancake and cucumber."* In Study 2B we added the details: *"Some laptops have a new feature: a* ***built-in LCD Projector****. The built-in projector does not impact the shape/form and weight of the laptop or the battery life."*

At the end of each survey, after participants completed the 12 choice tasks, we also collect demographics. Specifically, we collect typical survey demographic levels (options specified in parentheses): gender (male/female/non-binary/other:please specify), age, race and ethnicity (American Indian or Alaskan native/Asian, not including South Asian/South Asian/Black

or African American/Hispanic, Latino, or of Spanish descent/Native Hawaiian or Pacific Islander/White/other:please specify), political affiliation (Republican/Democrat/Independent/No preference/other:please specify), household income (Under $10,000/$10,000–$24,999/$25,000–$49,999,/$50,000–$74,999/$75,000–$99,999/$100,000–$150,000/over $150,000), and education level (Some High School or less/High School/Some college - not currently enrolled/Some college - currently enrolled/2 year Associate Degree/4-year Bachelor's Degree/-Master's Degree/Advanced Professional Degree(MD/JD) or Doctoral Degree (PhD)).

D.2 Human Sample Characteristics

Table D.1 reports demographics information for our human samples for our studies compared to recent US population studies.

Table D.1: Demographics compared to US population

	Participants	**US population**
Prop. Women	0.54	0.51
Median Age	35	38.9
Median Income	$50,000-$74,999	$75,149
Prop. HS Degree or higher	0.99	0.89
Prop. White	0.69	0.75
Prop. Black	0.1	0.14
Prop. Democrat	0.44	0.25
Prop. Republican	0.18	0.27

Notes: $N = 1,504$ participants across 5 studies. US demographic information is from the 2020 US Census Bureau data. https://www.census.gov/quickfacts/fact/table/US, accessed July 2024. US political affiliation is from Gallup, https://news.gallup.com/poll/15370/party-affiliation.aspx, accessed July 2024. Because we collected income brackets we are unable to calculate the median income but the median bracket is lower than the median income in the population.

E Prompts and Study Details

Below we provide the complete sets of prompts for our analyses and other study details. As mentioned in Online Appendix A, 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. Note that in GPT models 3.5-Turbo and higher for text completion, GPT expects a system message and a user message. The *system message* is a message from a developer to the GPT, explaining to it the expected behavior of the GPT in this context. The *user message* provides requests or comments for the GPT assistant to respond to. We run each

prompt as an independent query from any other prompt, and set temperature at 1.0 for all of our queries.

Prompts for our conjoint studies use variables that represent the different attribute levels, for example, in Study 2A the variables: $LaptopBrand1$,$Price1$, $RAM1$, $Storage1$, and $Size1$ correspond to the brand, price, RAM, Storage, and size for the first configuration in a consideration set.

E.1 Prompts for Pilot Studies (using the text-davinci-003 model)

E.1.1 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: "

E.1.2 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: $"

E.1.3 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 with/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: "

E.2 Prompt for Main Studies (using the gpt-3.5-turbo-0125 model)

E.2.1 Prompts for Study 1A

"**System message:**

You are a customer. You are selected at random while shopping for toothpaste to participate in a survey. The interviewer will describe the options you saw while shopping and ask you to report which option you chose to purchase. Whenever two options are shown, you can also choose a third option which is not to purchase anything that day.

User message:

While shopping, you see two options:

1. $ToothpasteBrand1$, Price: $\$Price1$, Has Fluoride: $Flouride1$, Flavor: $Flavor1$

2. $ToothpasteBrand2$, Price: $\$Price2$, Has Fluoride: $Flouride2$, Flavor: $Flavor2$

You also have the option not to purchase toothpaste.

Did you purchase either of the available toothpaste options? If so, which one?"

E.2.2 Prompts for Study 1B

For this study, the prompts are the based on the prompts of Study 1A, changes are illustrated in red.

"**System message:**

You are a customer. You are selected at random while shopping for toothpaste to participate in a survey. Some toothpastes have new flavors: pancake and cucumber. The interviewer will describe the options you saw while shopping and ask you to report which option you chose to purchase. Whenever two options are shown, you can also choose a third option which is not to purchase anything that day.

User message:

While shopping, you see two options:

1. $ToothpasteBrand1$, Price: $\$Price1$, Has Fluoride: $Flouride1$, Flavor: $Flavor1$

2. $ToothpasteBrand2$, Price: $\$Price2$, Has Fluoride: $Flouride2$, Flavor: $Flavor2$

You also have the option not to purchase toothpaste.

Did you purchase either of the available toothpaste options? If so, which one?"

E.2.3 Prompts for Study 2A

"**System message:**

You are a customer. You are selected at random while shopping for laptops to participate in a survey. The interviewer will describe the options you saw while shopping and ask you to report which option you chose to purchase. Whenever two options are shown, you can also choose a third option which is not to purchase anything that day.

User message:

While shopping, you see two options:

1. $LaptopBrand1$, Price: $\$Price1$, RAM: $RAM1$ GB, Storage: $Storage1$ GB, Size: $Size1$ in

2. $LaptopBrand2$, Price: $\$Price2$, RAM: $RAM2$ GB, Storage: $Storage2$ GB, Size: $Size2$ in

You also have the option not to purchase a laptop.

Did you purchase either of the available laptops? If so, which one?"

E.2.4 Prompts for Study 2B

For this study, the prompts are the based on the prompts of Study 2A, changes are illustrated in red.

"**System message:**

You are a customer. You are selected at random while shopping for laptops to participate in a survey. Some laptops have a new feature: a built-in LCD projector. The built-in projector does not impact the shape/form and weight of the laptop or the battery life. The interviewer will describe the options you saw while shopping and ask you to report which option you chose to purchase. Whenever two options are shown, you can also choose a third option which is not to purchase anything that day.

User message:

While shopping, you see two options:

1. $LaptopBrand1$, Price: $\$Price1$, RAM: $RAM1$ GB, Storage: $Storage1$ GB, Size: $Size1$ in, Built-in LCD Projector: $Projector1$.

2. $LaptopBrand2$, Price: $\$Price2$, RAM: $RAM2$ GB, Storage: $Storage2$ GB, Size: $Size2$ in, Built-in LCD Projector: $Projector2$.

You also have the option not to purchase a laptop.

Did you purchase either of the available laptops? If so, which one?"

E.2.5 Prompts for Study 2C

"**System message:**

You are a customer. You are selected at random while shopping for tablets to participate in a survey. The interviewer will describe the options you saw while shopping and ask you to report which option you chose to purchase. Whenever two options are shown, you can also choose a third option which is not to purchase anything that day.

User message:

While shopping, you see two options:

1. $TabletBrand1$, Price: $\$Price1$, RAM: $RAM1$ GB, Storage: $Storage1$ GB, Size: $Size1$ in

2. $TabletBrand2$, Price: $\$Price2$, RAM: $RAM2$ GB, Storage: $Storage2$ GB, Size: $Size2$ in

You also have the option not to purchase a tablet.

Did you purchase either of the available tablets? If so, which one?"

E.2.6 Prompts Including Demographic Characteristics

In these prompts, we modify the descriptor in the system message "You are a customer" as follows:

- Income: "You are a customer with an annual income of [less than $50,000/ $50,000–$74,999/ $75,000–$150,000/ over $150,000]. You are selected..."
- Gender: "You are a [male/female] customer. You are selected..."
- Political Affiliation: "You are a customer who is a [Republican/Democrat]. You are selected..."
- Race and Ethnicity: "You are a [While/Black] customer. You are selected..."
- Education: "You are a customer [without a college degree/with a college degree]. You are selected..."
- Age: "You are a customer, aged [28 or younger/ 28–42/ 42 or older]. You are selected..."

E.3 Number of observations collected

In the pilot studies, we collected 300 responses for each consideration set. In Studies 1A, 1B, 2A, 2B, and 2C we collected 150 responses for each consideration set. After exploring the consistency of these responses (see Online Appendix G), the rest of the studies were run with 50 responses for each consideration set, except for the heterogeneity studies on the fine-tuned GPT that were run with 30 responses for each consideration set for each group.

E.4 Costs

While our approach significantly reduces the cost and time it takes to collect survey data, using OpenAI fine-tuned models does become costly given current prices. Because we wanted to thoroughly study GPT's responses, we collected a large number of responses for each prompt, our studies included complex designs with many attributes and features, and we added runs that explored responses of different demographic groups one by one, all of which increased the cost of our exploration. For example, the cost of running Study 2A on GPT-3.5 Turbo was about $100, whereas adding one feature with two levels (with/without projector) in Study 2B increased the cost to about $500. However, we expect these costs to decline dramatically over time, and there are already multiple steps researchers can take to make our approach cheaper. When we learned that GPT responses converge quickly (see Online Appendix G) we were able to reduce the number of repetitions per prompt from 150 to 50. Moreover, OpenAI now offers batch queries which offer 50% lower costs for responses within 24 hours, which helped reduce the costs farther (i.e., the cost for Study 2B on a fine-tuned model were reduced from $2,600 to roughly $433 when applying both of these changes.), and new generations of LLMs continue to become cheaper to query and fine-tune.[19] While we expect that these costs will go down over time, as LLMs are commoditized, we also expect that firms that adopt our methodology will likely train and fine-tune an open-source LLM, which will significantly reduce costs.

F Sample GPT Responses

F.1 Responses for Study 1B

- "I chose the Colgate toothpaste, it has fluoride and mint flavor."
- "I would choose to purchase the Colgate toothpaste with the cinnamon flavor for $1.99 since it has fluoride, which is beneficial for my teeth."
- "I'd like to try the new flavor of Crest toothpaste with Cucumber. It sounds interesting and I don't mind the extra cost for the fluoride option."
- "I would like to try the new flavors. I choose not to purchase either of the available toothpaste options today."

[19] For example, after we completed our analysis, OpenAI announced their newest model, "GPT-4o mini", which is 60% cheaper per token than GPT-3.5 Turbo, and Meta introduced a 405B parameter open source LLM that is marketed as comparable to OpenAI's and Anthropic's advanced models (See https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ and https://ai.meta.com/blog/meta-llama-3-1/.)

F.2 Responses for Study 2A

- "I decided to purchase the Microsoft Surface Laptop with 32GB RAM, 256GB storage, and 15-inch size for $1000."
- "I chose to purchase the Apple MacBook Air."
- "I chose not to purchase either of the available laptops on that day."
- "I would choose Option 2: Apple MacBook Air, because it has more storage space compared to the Microsoft Surface Laptop."
- "I would go with option 1, the Apple MacBook Air. The specs meet my requirements and the price is more affordable compared to the Microsoft Surface Laptop."
- "I chose to purchase the Microsoft Surface Laptop as it had better specs for my needs, even though it was a bit pricier than the Apple MacBook Air."
- "I would choose the Microsoft Surface Laptop as it fits my budget better and has similar specifications to the Apple MacBook Air."

G Robustness and Additional Results

G.1 Consistency in GPT Estimates

Initially, we prompted GPT 150 times for each consideration set we wanted it to consider. Then, we randomly sampled the responses and estimated WTP to evaluate the consistency in its responses. Because of the costs of running GPT (especially the costs of querying the fine-tuned models), and given the consistency in the responses, we decided to reduce the number of repetitions to 50.

The table below reports the WTP estimates implied by Study 2B for different randomly selected sub samples of the data. Note that for Study 2B, given the number of attributes and levels, the total number of queries was: 3,483,000. This is because the total number of configurations is: (2 Brands X 3 Prices X 3 Storage X 3 RAM X 2 Size X 2 Projector), and the number of consideration sets of two products that include only options that differ from each other on at least one attribute based on n configurations is $\frac{n\times(n-1)}{2}$. We collected 150 responses per set. Note that due to the fact that some responses did not include a clear choice, we used 99.7% of GPT responses, and the total number of choices used in estimation for this study was 3,473,621. We

also note that to produce the table below, the standard error calculations for the multinomial logit (as if the data was from a random sample of consumers) imply tight confidence intervals even for the 10% sample, and for most of the attributes in the 1% sample.

Table G.1: WTP Estimates based on GPT Study 2B Responses

	100%	**50%**	**25%**	**10%**	**1%**
Surface	922.83	918.90	926.33	927.83	868.64
Macbook	819.53	815.94	823.70	824.95	764.97
RAM	14.50	14.51	14.49	14.51	14.41
Storage	0.48	0.48	0.47	0.48	0.48
Size	-12.48	-12.27	-12.64	-13.21	-9.84
Projector	1074.31	1076.26	1070.69	1081.14	1074.88

Notes: Each Column reports the WTP estimates based on a randomly selected sample of different size, where Column 1 uses 100% of the sample, and Column 2 uses 50% of the sample.

G.2 Comparing GPT and Humans Across Demographic Groups

Table G.2 reports the comparison of WTP and no purchase rates from humans and GPT for Study 2A, and Table G.3 reports the same for Study 2B. For Study 2A, we simply report WTP estimates based on our human sample and the responses from GPT. For Study 2B, we present the results from GPT as is, and from a fine-tuned version of GPT, where we fine-tune GPT-3.5 Turbo with the Study 2A responses, but we modified the prompt to include the income and gender of the customer, based on our survey responses. For example, if the survey respondent was a female with income bracket $75,000—$150,000 the training prompt was adapted to say "You are a female customer with an annual income of $75,000—$150,000. You are selected...". Out of the 302 responses for Study 2A, we had 6 cases with missing income or gender, and so we omitted them from the fine-tuning data. We then queried the fine-tuned GPT with Study 2B questions where we separately ask about different income level, gender, and political affiliation.

Table G.2: WTP Estimates from Study 2A for Human and GPT Demographics Groups

Group		Sample	RAM	Storage	Size	MacBook	Opt out
			Income				
Human	< $50K	35%	17.8	0.63	69.6	12.5	49%
Human	$50-74,999	21%	26.5	0.82	82.2	32.2	50%
Human	$75-150K	29%	22.1	0.47	68.3	51.1	52%
Human	> $150K	13%	19.4	0.48	47.8	225.6	47%
GPT	< $50K		-4.4	0.05	-27.3	-2.6	95%
GPT	$50-74,999		2.5	0.19	-27.1	39.0	34%
GPT	$75-150K		22.9	0.61	1.2ˆ	176.0	19%
GPT	> $150K		99.6	2.31	126.4	722.5	34%
			Gender				
Human	MALE	44%	22.9	0.63	87.2	-46.9	55%
Human	FEMALE	54%	19.1	0.57	47.8	164.9	46%
GPT	MALE		15.2	0.52	-1.5	17.3	9%
GPT	FEMALE		13.5	0.53	-27.3	113.8	26%
			Political Affiliation				
Human	REPUBLICAN	15%	16.2	0.57	80.8	41.7	53%
Human	DEMOCRAT	52%	19.5	0.52	57.7	113.4	50%
GPT	REPUBLICAN		5.5	0.26	1.2ˆ	-252.0	27%
GPT	DEMOCRAT		16.2	0.37	-24.5	373.4	41%
			Race and Ethnicity				
Human	WHITE	70%	25.8	0.79	94.8	60.1	54%
Human	BLACK	10%	15.9	0.28	58.3	-52.1	37%
GPT	WHITE		13.7	0.52	-26.1	315.1	21%
GPT	BLACK		15.4	0.54	-1.4ˆ	-135.5	65%
			Education				
Human	NO COLLEGE	39%	21.0	0.68	84.9	-61.4	48%
Human	ONLY COLLEGE	44%	20.5	0.57	55.0	128.4	55%
GPT	NO COLLEGE		-0.5	0.18	-25.9	-48.6	55%
GPT	COLLEGE		21.4	0.67	-25.8	-92.9	46%
			Age				
Human	AGE<28	25%	20.7	0.43	23.5ˆ	77.6ˆ	52%
Human	[28-42]	50%	19.4	0.70	84.9	92.5	45%
Human	AGE>42	25%	25.7	0.56	86.6	-40.5	59%
GPT	AGE<28		9.0	0.36	-40.3	21.6	31%
GPT	[28-42]		15.0	0.53	-8.4	39.3	11%
GPT	AGE>42		19.2	0.76	1.5ˆ	74.7	29%

Notes: For our comparison, we use human groups which make up at least 10% of the human sample. The Column "Sample" indicates what fraction of the human sample corresponds with this group. Column "Macbook" indicates the preference for Macbook (over Surface). Column "Opt out" indicates the no purchase shares." The Human "No College" sample includes all of those who didn't start or complete college. Age splits are based on quartiles. For the GPT samples, 50 responses were collected for each group. See prompts in Online Appendix E. The symbol ˆ indicates cases where the confidence interval in the multinomial logit for a specific attribute included zero.

Table G.3: WTP Estimates from Study 2B for Human and GPT Demographics Groups

Group		Sample	RAM	Storage	Size	MacBook	Projector	Opt out
				Income				
Human	< $50K	30%	18.0	0.39	74.9	-52.7	219.1	48%
Human	$50-74,999	22%	23.7	0.62	127.2	119.7	423.5	35%
Human	$75-150K	34%	19.3	0.77	29.5	212.1	430.0	41%
Human	> $150K	11%	17.0	0.44	58.4	203.7	142.0	45%
GPT	< $50K		-2.2	0.09	-28.5	-27.9	-118.2	85%
GPT	$50-74,999		4.8	0.23	-24.8	-35.1	142.9	37%
GPT	$75-150K		14.5	0.47	-15.9	-70.0	636.8	24%
GPT	> $150K		103.9	2.44	73.9	-182.9	2786.6	48%
GPT-FT	< $50K		20.5	0.47	26.6	46.7	169.4	48%
GPT-FT	$50-74,999		21.6	0.48	26.1	58.9	169.4	49%
GPT-FT	$75-150K		23.5	0.52	29.7	52.1	173.6	48%
GPT-FT	> $150K		25.1	0.54	28.7	60.5	191.7	48%
				Gender				
Human	MALE	43%	19.9	0.47	65.3	-50.2	288.3	43%
Human	FEMALE	56%	18.2	0.64	65.6	249.2	350.9	42%
GPT	MALE		11.1	0.37	-14.8	-74.2	756.4	11%
GPT	FEMALE		11.5	0.40	-23.0	-73.4	976.8	18%
GPT FT	MALE		21.7	0.51	29.6	45.2	176.3	49%
GPT FT	FEMALE		20.9	0.49	16.7	56.7	170.8	44%
				Political Affiliation				
Human	REPUBLICAN	23%	16.9	0.57	28.9	-101.2	255.1	41%
Human	DEMOCRAT	40%	15.8	0.44	52.5	308.9	308.4	39%
GPT	REPUBLICAN		3.5	0.15	-6.8	-133.5	-57.6	24%
GPT	DEMOCRAT		10.0	0.20	-23.8	-29.9	817.5	41%
GPT-FT	REPUBLICAN		22.7	0.49	22.6	67.6	195.9	47%
GPT-FT	DEMOCRAT		23.1	0.50	31.3	52.1	176.9	48%

Notes: For our comparison, we use human groups which make up at least 10% of the human sample. The Column "Sample" indicates what fraction of the human sample corresponds with this group. Column "Macbook" indicates the preference for Macbook (over Surface). Column "Opt out" indicates the no purchase shares." For the GPT-3.5 Turbo samples, 50 responses were collected for each group, for the fine tuned model (GPT-FT in the table), 30 responses were collected for each group. See prompts in Online Appendix E.