BVB BibliothekenVerbund Bayern

Aufsatzbestellung von 1049 - FERNLEIHE

Signatur: 4 Z 82.26 / Hbzs 602-36 = Neueste Hefte * 2024, oA

URL:

Bestellnummer: **20240058637**

Bestelldatum: 29.02.2024 / Eingangsdatum: 29.02.2024

Lieferart: **PRINT**

Kostenübernahme: **0 Euro**

Bestellinstitution: **[1049] Hochschule für angewandte Wissenschaften Neu-Ulm, Hochschulbibliothek (flnehmend@hnu.de) Postfach 17 44, 89207 Neu-Ulm**

Benutzer: **Jordan, Stephanie (NU950005682)**

E-Pub ahead of print

Benachrichtigung: E-Mail an Besteller

Lieferinstitution: [12] Bayerische Staatsbibliothek (bvbafl@bsb-muenchen.de)

Buch/Zeitschrift: **Marketing science**

Autor/Hrsg.:

ISSN/ISBN: 0732-2399

Ort, Verlag: Catonsville, MD, INFORMS

Titel des Aufsatzes: Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis

Autor des Aufsatzes: Li, Peiyao

Jahrgang: 2024

Band:

Heft:

Seitenangabe: oA

Gesamttitel:

Sammelfeld: EF:27.08.2024 AO:00/01 Fernleihstelle - Ausleihtheke

Hinweise: InfoGuide: default: PrimoCentral. - DOI10.1287/mksc.2023.0454

Weiterleitungen nach: [1.] 945, [2.] 355, [3.] 703, [4.] HBZ, [5.] GBV, [6.] KOBV

Jordan, Stephanie (NU950005682)

informs.
https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Articles in Advance, pp. 1–13
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis

Pelyao Li,[a] Noah Castelo,[b] Zsolt Katona,[a,*] Miklos Sarvary[c]

[a] Haas School of Business, University of California, Berkeley, California 94720; [b] Alberta School of Business, University of Alberta, Edmonton, Alberta T6G 2R6, Canada; [c] Columbia Business School, Columbia University, New York, New York 10027

*Corresponding author

Contact: ojhfklsjhl@berkeley.edu (PL); ncastelo@ualberta.ca (NC); zskatona@haas.berkeley.edu, https://orcid.org/0000-0001-8411-6040 (ZK); miklos.sarvary@columbia.edu, https://orcid.org/0000-0002-3301-5917 (MS)

Received: September 28, 2023
Revised: November 7, 2023
Accepted: December 5, 2023
Published Online in Articles in Advance: January 25, 2024

https://doi.org/10.1287/mksc.2023.0454

Abstract. This paper explores the potential of large language models (LLMs) to substitute for human participants in market research. Such LLMs can be used to generate text given a prompt. We argue that perceptual analysis is a particularly promising use case for such automated market research for certain product categories. The proposed new method generates outputs that closely match those generated from human surveys: agreement rates between human- and LLM- generated data sets reach over 75%. Moreover, this applies for perceptual analysis based on both brand similarity measures and product attribute ratings. The paper demonstrates that, for some categories, this new method of fully or partially automated market research will increase the efficiency of market research by meaningfully speeding up the process and potentially reducing the cost. Further results also suggest that with an ever larger training corpus applied to large language models, LLM-based market research will be applicable to answer more nuanced questions based on demographic variables or contextual variation that would be prohibitively expensive or infeasible with human respondents.

History: Catherine Tucker served as the senior editor. This paper was accepted through the *Marketing Science*: Frontiers review process.

Funding: This work was supported by the Social Sciences and Humanities Research Council of Canada [Grant 430-2021-00057].

Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mksc.2023.0454.

Keywords: artificial intelligence • perceptual maps • large language model • natural language processing • market research

1. Introduction

Language analysis has become an integral part of marketing research. Various tools and methods have been developed to automate the text analysis process once the data are generated and collected (Berger et al. 2022). We explore the potential of large language models (LLMs) to generate data for the purpose of producing market research, thereby improving on or even eliminating the need for surveys and human respondents. LLMs are trained on large bodies of text collected from diverse sources with the primary purpose of mimicking human written text. For example, generative pretrained transformer (GPT) models predict the next word after a prompt (Radford and Narasimhan 2018). Such transformers use the self-attention mechanism developed in Vaswani et al. (2017) to effectively capture the context in the text and are often combined with reinforcement learning with human feedback (RLHF) to better mimic human dialogue (Ouyang et al. 2022).

The goal of this paper is to assess whether and when appropriate adaptations of these models can substitute for market research using human surveys. The practical benefit of such substitution is the potential to significantly scale up market research activity. Moreover, LLM-based market research may be able to dig deeper and answer more complex questions or uncover information about hard-to-reach demographic segments. We first ask: How well does market research based on LLMs replicate market research based on human respondents for generating specific consumer insights that are useful for marketers?

We focus on one particular market research task: perceptual analysis. Marketers have long been interested in mapping how consumers perceive brands. The primary source of data for analyzing perceptions are surveys where respondents rate how similar brands are and how they score on different attributes. Conducting representative surveys requires a large number of participants, which is difficult to do at scale regularly and in many different contexts. A well-calibrated language model adaptation can collect the required data automatically. Our goal is to demonstrate a "use case" of such LLM-based market research.

Importantly, when it comes to measuring brand *perceptions*, the ultimate source of data are consumer self-reports. This is not necessarily the case for other marketing variables; for example, preferences, where sales and market shares (i.e. consumer choice) may be more accurate measures of, *revealed* preferences than consumers surveys. When it comes to brand perceptions, consumers are typically asked about the associations and meanings that brands represent to them. Because language models have been engineered to accurately replicate human responses, they are particularly adept tools for perceptual analysis (see Horton (2023) and Brand et al. (2023) for similar arguments).

The core of a large (generative) language model takes a prompt as an input and generates a continuation of text as output. Formulating the best prompt and analyzing the generated output presents a number of challenges. In our context, a typical human brand perception survey would directly ask people how similar two brands are: "How similar are the brands BMW and Audi?" The respondent would be expected to give a numerical answer, say between 0 and 10. With the right prompt, some of the most advanced models at the time of writing the paper, like GPT-4, can be asked to do just that.[1] However, it is often difficult to assess these numerical answers' reliability and internal consistency (Espejel et al. 2023). To benchmark their performance, besides similarity ratings we also use an open-ended prompt with only one brand mentioned: "The car brand BMW is similar to the car brand ….," where the model fills in the blank. We then search for the mentions of other brands in the output, such as "Audi" or others. After rigorous analysis of the frequencies of brands appearing after repeated prompting, we can generate similarity scores between pairs of brands.

The paper's main objective is to demonstrate the validity of this newly developed LLM-powered data collection and analysis methodology. For validation, (i) we collect human survey responses, and (ii) we compare against car trade-in data capturing consumer switching behaviors between brands. In both cases, we demonstrate that the artificially generated responses and the resulting perceptual structure closely resembles what the human data yield. Comparing the output of data generated by language models and human surveys is not trivial. We develop a novel methodology, called the "triplet method" to address several challenges in comparing the various data sets. The obstacles presented by the differing formats and variable baseline effects require a generic method that relies on the order of similarities between three objects (three brands or a brand and a pair of attributes). We demonstrate that this new method is a robust way to compare perceptual data sets generated by machines and humans and show that agreement rates between them can reach more than 75% for some product categories.

We examine two types of perceptual analysis: one relying on brand similarities and another based on the most important attributes and features that contribute to a brand's perceived image. The first application is typically used in high-level, top-down analysis of perceptions. We use EvoMap developed by Matthe et al. (2022) and *t*-distributed stochastic neighborhood embedding (t-SNE) to create corresponding perceptual maps. The second type of perceptual analysis we conduct is a bottom-up, deeper analysis that is based on consumers' ratings of brands on category-specific attributes (e.g., safety, economy, sportiness, etc., for cars). Factor analysis of these ratings reveal fundamental components of category brand perception. Again, our results show human and LLM responses generating highly similarity perceptual maps.

LLMs are highly customizable through prompting. We expand our method by exploiting prompts to account for the (i) time of the inquiry and (ii) the demographic characteristics of the respondent. The first task is accomplished by injecting a designated year into the prompt. We then compare the LLM results with trade-in data for the given year and find that our method is able to account for time in this fashion. Second, we explore heterogeneity in consumer responses based on demographics. Importantly, perceptual analysis, if representative, generally does *not* exhibit consumer heterogeneity. For example, a car brand (say, Ferrari) only works as a consumer signaling device if *all* consumers perceive it roughly the same way, including those consumers who are not in the target segment. Indeed, Ferrari is generally recognized as a "sporty" car even by consumers who would never consider purchasing it. In contrast, brand *preferences* should significantly differ across customer segments. For this second task, we modify the prompts to account for individual characteristics, such as age, gender, and income; variables that we can observe in the human survey data. We then compare the results across the different groups within the LLM-generated responses and between the human and LLM responses. As expected, we find no significant differences in *perceptions* across segments neither in the human nor in the LLM-generated data. However, the differences in *preferences* are significant across demographic segments and they are highly consistent across human and LLM data.

This extended analysis illustrates how the general methodology advanced in this paper can be adapted to specific circumstances. Finding a sample of human respondents who are representative of a certain narrow demographic group or learning how perceptions change over time is highly challenging if not impossible for most firms. One advantage of using algorithmic responses is the ease of interrogating the model about the beliefs of even narrowly defined demographic groups at specifically determined times or contexts.

Our approach follows in the path of an emerging literature in marketing that uses user-generated content (UGC) to infer competitive market structures in specific product categories. The objective is to provide firms with a cost efficient and rapid way to gain insights in their competitive position as is represented in consumers' minds. A pioneering example is Netzer et al. (2012), who apply text-mining techniques and semantic network analysis to comments on a public forum to deduce the market structure of sedan cars and diabetes drugs. Tirunillai and Tellis (2014) provide a different method applied to the same problem, which allows them to extract a more refined view of consumer perceptions, such as the underlying dimensions of a particular concept (e.g., quality) or consumer heterogeneity. Using a much broader set of (200) brands, Culotta and Cutler (2016) use UGC data on Twitter to infer attribute-specific brand perceptions. They also compare the method's automatic ratings estimates with directly elicited survey data and find a consistently strong correlation. Humphreys and Wang (2018) provide a quite extensive review of this literature and explore the broad opportunities and limits of using text analysis for consumer research. Timoshenko and Hauser (2019) show that user-generated content is valuable for evaluating customers' needs, and they develop a machine-learning based approach to make identifying these needs from user-generated content more efficient. Dzyabura and Peres (2021) introduce a machine-learning based image processing method that allows firms to extract objects or emotions associated with different brands and then use these meaningful associations to gain a competitive advantage. Liu et al. (2020) develop a deep-learning model that predicts perceptual brand attributes using online images posted by consumers.

More recent papers have expanded the scope of data available on the Web to develop consumer insights. Liu and Toubia (2018) use an extended latent Dirichlet allocation to relate topics in search queries to the topics in the corresponding top search results and estimate consumers' content preferences from search queries on a search page. Matthe et al. (2022) use similarity data derived from company 10-k Securities and Exchange Commission filings to estimate market structure evolution over time.

Our research also relies on data broadly available on the Web, and we aim to understand market structures by estimating consumer perceptions without directly eliciting these perceptions in surveys. A departure from previous methods is that we use a large language model to "collect" the relevant data or to define the desirable product context. As mentioned earlier, these models aim at replicating human responses, and their training sets encompass a vast representation of human "thinking."

Overall, our main contribution is opening the door to a new way of conducting market research with no or far fewer human respondents. We develop new methodology and a workflow that allows a marketer to rely only on an LLM to conduct market research. Although we use two specific language models (GPTNeo 2.7B and ChatGPT based on GPT4), our method can be adapted to any large (generative) language model. With the high pace of development in this area, we expect these models to become more accurate and faster to run. We demonstrate that even at the current state of models, our LLM-powered market research method can produce meaningful results and replicate human results. Furthermore, we explore some of the ways that algorithmic market research can surpass the capabilities of research conducted with human participants.

2. Methods

Our goal is to use LLMs to perform perceptual analysis. Figure 1 provides the roadmap of our methodology. We describe each step in the roadmap and highlight how an LLM may assist marketers to perform them.

2.1. Steps 1–3: Defining the Set of Relevant Brands and Attributes

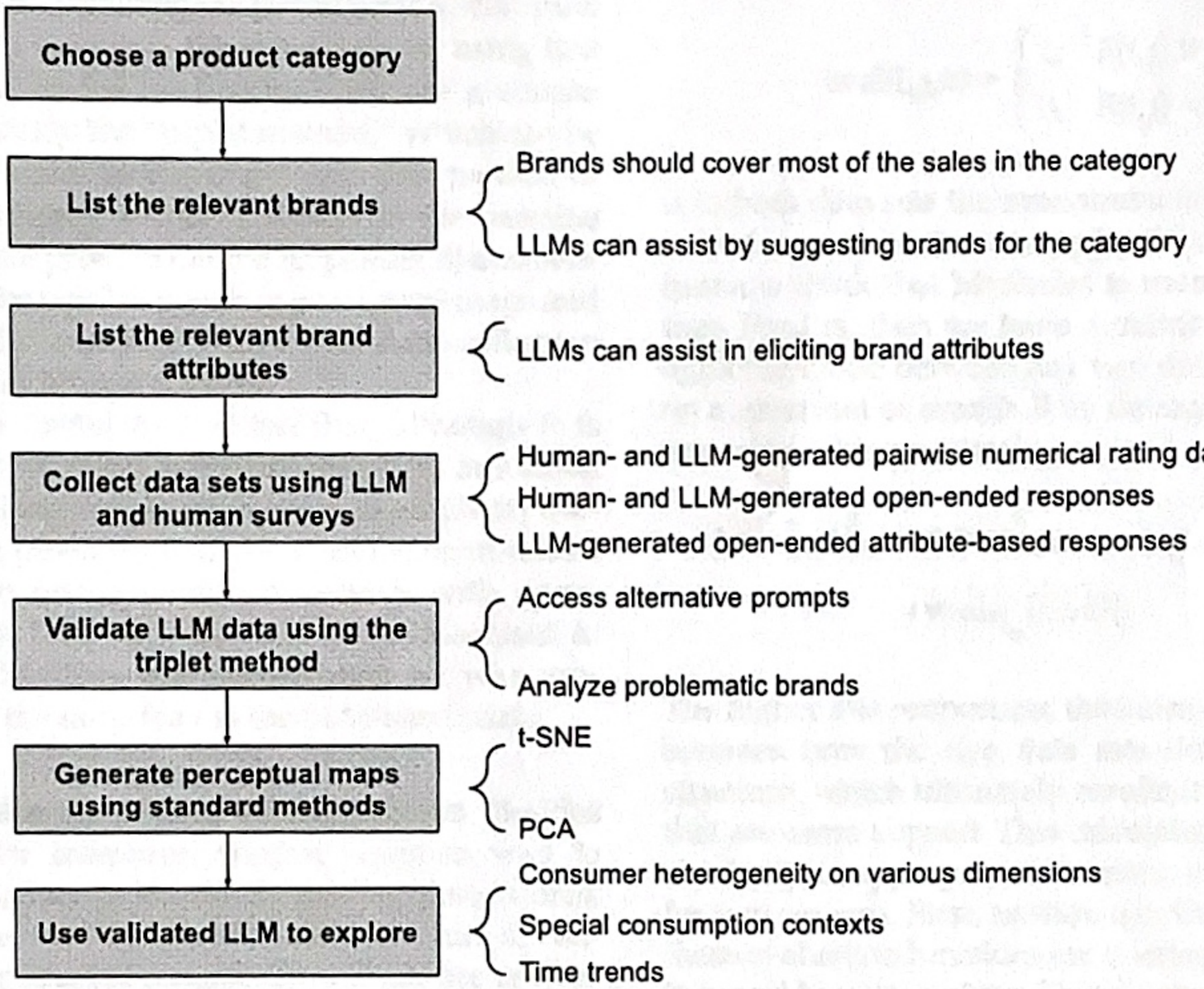
Perceptual analysis always starts with the selection of a product category and the brands that are representative of the category. These steps always rely on a number of judgment calls that depend on the marketer's objectives (e.g., how narrow/broad should the category be). Similarly, which brands should represent the category may vary by study (e.g. the geography studied), although the set should generally cover a critical proportion of the category sales and should include all brands beyond a critical market share. Although LLMs can contribute to the generation of the brand set, here their added contribution is limited. For example, in our application for the car category, we used the LLM to generate the brand set by picking the 21 brands with the highest frequency when the LLM was asked to complete the prompt "The car brand ...." Our cutoff at 21 brands was a judgment call.

If the marketer chooses to use attribute-based perceptual analysis, the relevant attribute space needs to be defined for the category. A product category (e.g. cars) is typically described by a list of attributes ("comfortable," "sporty," etc.) and associated scores along each attribute for each brand. Again, LLMs can provide (limited) assistance in this process by eliciting an initial list of brand attributes (see Online Appendices C and D for details on the associated prompts). Importantly, as in the case of brand selection, a fair amount of judgment needs to be applied to finalize the list of attributes used.

2.2. Step 4: Collecting LLM- and Survey-Based Data

Perceptual analysis can be based on two types of data: (i) brand similarity measures and (ii) brand attribute

Figure 1. (Color online) Roadmap for LLM-Based Perceptual Analysis



scores. It is generally beneficial to perform both analyses for more robust insights.

2.2.1. Data to Measure Brand Similarity. The most basic way to draw perceptual maps involves asking people in a survey for direct numerical values of pairwise similarity ratings between two brands. Then, the mean value of all numerical ratings for a given pair of brands provides the brand similarity measure. Using the appropriate prompt, certain LLMs are capable of replicating this method by providing numerical answers with a reasonable level of reliability.[2]

LLMs can also generate similarity measures using open-ended text data, which constitutes a more "natural" use of these models. Consider the category of car brands. When we prompt the LLM with "The car brand BMW is similar to the car brand ...," one of the LLM-generated continuations is "Ford. They are both car manufacturers." We consider this LLM-generated text as the raw input data to the analysis. Specifically, from such continuations, we can build a frequency matrix with cells corresponding to each brand pair. The (i, j)th entry of this frequency matrix is the number of times brand i appeared after a prompt asking about brands similar to j. For example, if the word "BMW" appears a total of 10 times in all the completions we collected for the prompt "The car brand Audi is similar to the car brand ...," we record the (BMW, Audi) cell in the frequency matrix as 10.

However, these frequencies can be biased by brands' top-of-mind awareness or other brand-specific features that affect the overall mentions of the brand. We call this bias the baseline value of a brand. A high baseline brand like Ford may be mentioned frequently because it is a well-known car brand and not because it is actually similar to the brand given in the prompt. To explicitly address this baseline bias, we apply an ordinal embedding method.

As discussed in Jain et al. (2016), an ordinal embedding method can estimate similarity scores of objects given their qualitative relative distances like "A is closer to B than to C."[3] For a comparison of brand A relative to B and C, we count whether the brand name B or C appears more often following the prompt containing brand A. To generate a similarity *score* between a *pair* of brands, we compare each of these two brands with all other brands. Take the case of four brands: BMW, Mercedes, Audi, and Ford. We evaluate the similarity score between BMW and Mercedes by assessing whether both are more similar to Audi than to Ford. If this holds true, we conclude that BMW and Mercedes are similar relative to Audi and Ford. In practice, a larger number of brands is required. With more than four brands, the similarity score between BMW and Mercedes becomes

a proportion reflecting their similarity relative to any other brand pair. This method is designed to compute the similarity of brands in the same data set using text frequency data. In the next section, we use a similar intuition to develop the "triplet method," which can be used to evaluate similarities of *data sets*. This method of generating similarity scores is robust to the baseline effect (see formal proof in Online Appendix B) and subsequently can be used to generate perceptual maps and to meaningfully compare with other data collection methods such as human surveys.

Finally, it is useful to mention that, although it is more practical to collect similarity measures in human surveys by asking for direct numerical similarity ratings, to have a direct comparison with the open-ended LLM data, we also ran human surveys with open-ended prompts that were identical to the ones used to queue the LLM. This raw human data set was processed exactly the same way as the LLM-based data.

2.2.2. Data to Measure Brand Attribute Scores. Besides using similarity measures, another common way to draw perceptual maps is from brands' attribute scores. Here, attribute-level drivers of market structure in conjunction with principal component analysis are used to uncover the main features that define a given market.

LLMs can be efficiently used to generate brand attribute scores. To do so, we use the prompt: "The most comfortable car brand is ...." The frequency of a brand appearing in the output (i.e., the continuation text) is indicative of the brand's attribute score. In the corresponding human data collection task, we directly ask human respondents to rate brands on the set of attributes defined in Step 3. We discuss the details of this attribute-based perceptual analysis in Online Appendix C.

2.3. Step 5: Compare Data Sets for Validation

To compare the LLM-generated perceptual data to those we obtained from the human surveys, we develop a novel, nonparametric, data set-level comparison technique that we call the *triplet method*. In our basic analysis, we have four types of brand-similarity data sets, each based on a different data generation process described in Section 2.2: (i) LLM open-ended responses, (ii) LLM pairwise numerical ratings, (iii) human open-ended responses, and (iv) human pairwise numerical ratings.

The triplet method calculates data set level agreement rates between these different data sets as follows.[4]

For a triplet of brands (i, j, k), we assess whether brand j or k is more similar to i, separately in each data set. Mathematically, for a set of brands B and a data set D, we use the function $F_D^B(i,j)$ to denote the observed similarity measure between brands $i, j \in B$. The F function is based on observed frequencies in the open-ended data and on pairwise mean ratings in the pairwise numerical data. We define the triplet evaluation function as

$$\text{eval}(\mathbf{i},j,k) = \begin{cases} 1, & F(i,j) > F(i,k) \\ 0, & F(i,j) < F(i,k). \end{cases}$$

If in both data sets the assessment is identical, then we record a match. For example, if both the LLM and humans think that Mercedes is more similar to BMW than Ford is, then we have a match. We calculate the agreement rate between any two data sets $F_{D_1}^B$ and $F_{D_2}^B$ on a given set of brands B by determining the proportion of matching triplets[5]:

$$\text{AR}(F_{D_1}^B, F_{D_2}^B) = \text{mean}_{i,j,k \in B, i \neq j, i \neq k, j \neq k} \mathbb{1}[\text{eval}_{F_{D_1}^B}(i,j,k) = \text{eval}_{F_{D_2}^B}(i,j,k)].$$

The higher this proportion, the more agreement there is between how the two data sets describe the market structure, which ultimately results in perceptual maps that are more aligned. This calculation of data set level similarity is appropriate for many marketing settings for two reasons. First, as shown in Online Appendix B, these evaluation functions are unaffected by differences in brand baseline values. This allows us to analyze frequency tables converted from open-ended responses. Second, when a consumer considers switching brands and they are forming a consideration set, they may not directly consider the exact value comparison of the brands; rather, they would rely on a binary evaluation of whether brand i is better than brand j. Our evaluation function thus reduces the comparison problem to a binary problem: Is brand i more similar to j or to brand k?

The triplet method also allows us to calculate a theoretical maximum agreement rate for the observed agreement rates. To establish this maximum, we compute the self-consistency rate for each data set. For a data set with n data points, we create two equal-sized data sets by resampling the original data set with replacement. The self-consistency rate is the agreement rate between these samples.

When resampling from both data sets we are able to analytically show that in expectation there is an upper bound of the agreement rate using the self-consistency rates. This upper bound between any two data sets is the average of the two self-consistency rates. Therefore, when we try to compare the distribution of LLM-generated responses and human responses, we use the average of the self-consistency rates as a theoretical maximum.

2.4. Steps 6 and 7: Using the Model, Perceptual Maps, and Exploration

An important part of perceptual analysis is visualization, that is, the generation of perceptual maps. These

can provide qualitative insights to managers, which in turn can inform a variety of marketing actions from product positioning to research and development (R&D) decisions. From the data generated in Step 4, it is easy to draw perceptual maps using standard techniques. When using overall similarities of brands, we directly apply EvoMap's t-SNE implementation to the similarities. We take advantage of EvoMap's ability to align projections from high dimensional data sets. With a low to moderate level selected for the smoothing parameter, we can denoise the maps while preserving a relatively high amount of variation among different maps. We choose t-SNE because it can preserve the local structure of the underlying data. When using attribute-based similarities, we conduct preprocessing on the data and apply Principal Component Analysis to the processed data (please refer to Online Appendices C and D for a more detailed description of the preprocessing steps).

A key advantage of our proposed LLM-based method is that additional data can be easily generated by appropriately modifying the prompt. As such, once validated, the model can be efficiently used to explore a variety of scenarios and context. In our applications presented below, we show how appropriate prompt engineering can provide insights with respect to consumer heterogeneity and the evolution of perceptions over time.

3. Results of the Study of Car Brands

We first apply our methodology to a set of 21 major car brands. The list was assembled by ranking the brands with the highest frequency when the LLM is asked to complete the prompt "The car brand ...." Overall, these car brands contributed to more than 75% of the total number of cars sold in the United States in 2021.

3.1. Comparison with Survey Data

We collect 17,000 data points for each listed car brand by queuing the LLM with prompts "The car brand X is similar to the car brand ...." With the same prompts, we gather 402 human responses. LLM responses are 10 to 20 words long, and more than 95% of human responses are under three words, often single-brand mentions. Using this text data, we use bag-of-words ordinal embedding for brand pair similarity scores. Additionally, we collect 5,300 direct pairwise similarity ratings from 530 participants, yielding over 20 human numerical ratings for each brand pair. To compare, we gather GPT4 numerical similarity ratings for each car brand pair.

In Figure 2, we present perceptual maps where EvoMap's t-SNE–based implementation is applied to LLM-generated and human *open-ended* text transformed by ordinal embedding and LLM and human generated direct rating data. From left to right, these maps are constructed based on LLM open-ended data generated by GPTNeo, human open-ended survey data, LLM direct rating data generated by GPT4, and human direct rating survey data.

To more objectively examine the consistency between LLM and human results, we compute the agreement rate between LLM and human data using the nonparametric triplet method described in Section 2.3. To establish the theoretical maximum of our comparison, we compute the self-consistency rate of each data set by computing the agreement rate of two bootstrapped samples of the same data set. We repeat this bootstrap process to construct confidence intervals.

The agreement rates highlighted in Figure 3 are adjusted for the theoretical maximum. On the *open-ended* responses the agreement rate between the human-LLM data sets is 80.1% of the theoretical maximum of 93.6%. When comparing LLM *pairwise* numerical rating responses to human *pairwise* numerical ratings, we observe an agreement rate 87.2% as high as the theoretical maximum of 94.7%. When comparing open-ended responses versus direct rating responses, we see a lower agreement rate between LLM and human responses, but these adjusted agreement rates are still more than 70%. It is noteworthy that comparing the human responses that originate from the open-ended survey with those from the pairwise data yield a

Figure 2. (Color online) Perceptual Maps Using Overall Similarity Score and EvoMap (t-SNE)

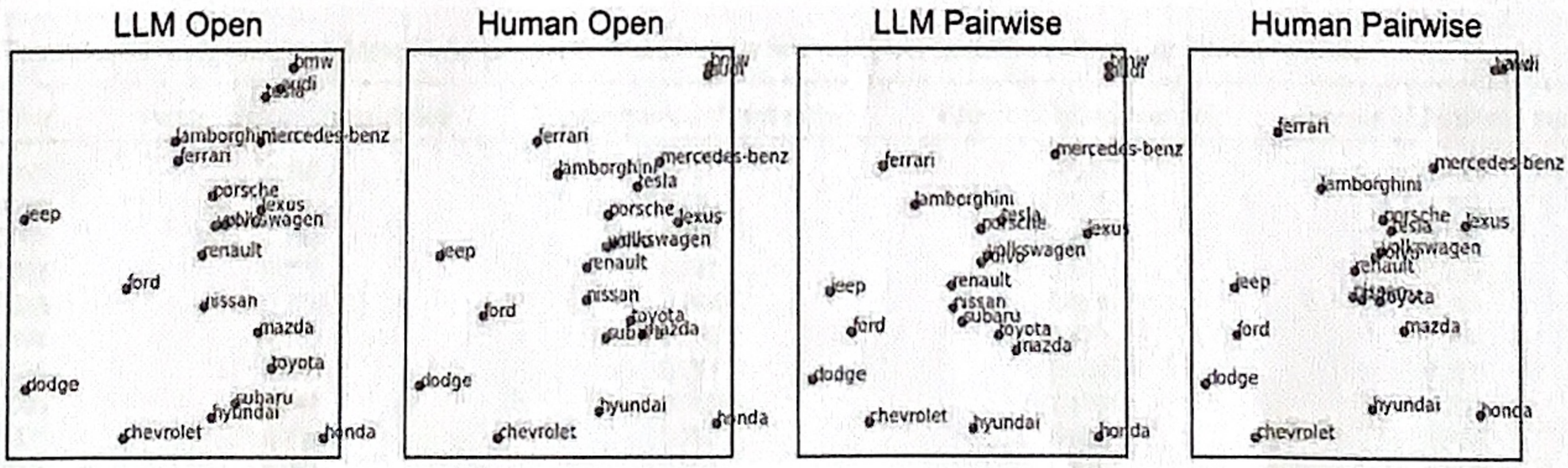
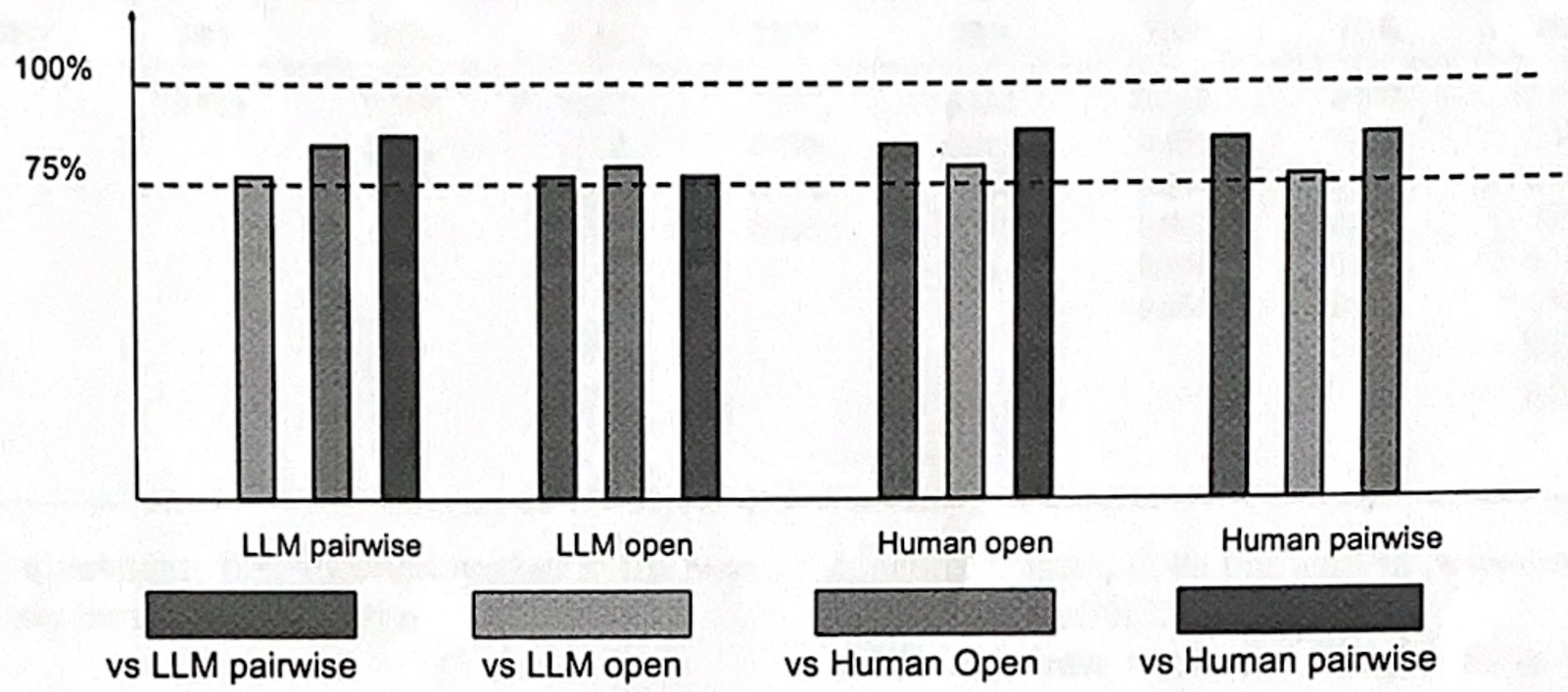


Figure 3. (Color online) Adjusted Agreement Rates of Car Brands in the Overall Similarity Study



surprisingly low adjusted agreement rate of 89.3%. That is only slightly higher than the 87.2% agreement rate between LLM-based and human pairwise responses.[6]

3.2. Comparison with Real Consumer Purchase Data

We have demonstrated LLM models' ability to produce brand perception data similar to human survey responses. Yet, survey answers might not align with actual consumer choices. To delve further, we match LLM-generated brand similarity data with real car switching data spanning 1999 to 2008.[7] This data set comprises quarterly observations of car trade-ins in the United States, encompassing 785 car models across 46 brands. These brands represent more than 97% of the entire U.S. automobile industry market share during this time frame. We treat the number of times consumers switched to brand A from B as a measure of the response to the question "how similar are car brands A and B."[8] Because the car trade-in data set contains brands not in our focal set, we scale these ratings by the total number of trade-ins from brand B to all brands in our focal set. We then apply our triplet method to compare this data set with the other human and LLM-generated data sets, calculating the agreement rates using the same method as before.

Table 1 shows raw agreement rates between the trade-in data and our four datasets (human direct rating, human open-ended, LLM direct rating, and LLM open-ended). LLM open-ended data exhibits 60%–70% raw agreement rates, whereas LLM direct rating and human data show 70%–80% rates. Raw agreement rates are akin to those between LLM and human survey data, signifying strong similarity of LLM-generated brand data to human surveys and actual consumer behavior.

We also observe a significantly positive correlation between the recency of the trade-in data and the data's agreement rates with both human and LLM survey data; that is, both LLM and human data are more similar to more recent car trade-in data. Temporal progression potentially alters consumer perceptions of car brands. Table 2 confirms the changing car trade-in patterns over time: each extra year corresponds to a significant 1.16% decline in agreement rates for year-specific consumer trade-in data. A natural question is whether we can extract year specific information from the language model through specifying the time period in the

Table 1. Raw Agreement Rates Between LLM and Human Survey Data and Consumer Car Trade-in Data

Year	Versus human open-ended	Versus human direct rating	Versus LLM open-ended	Versus LLM direct rating
1999	0.762	0.729	0.665	0.725
2000	0.764	0.736	0.672	0.738
2001	0.766	0.735	0.664	0.736
2002	0.772	0.741	0.669	0.739
2003	0.774	0.740	0.669	0.743
2004	0.785	0.758	0.672	0.756
2005	0.783	0.757	0.669	0.754
2006	0.787	0.765	0.674	0.751
2007	0.788	0.767	0.687	0.754
2008	0.796	0.774	0.689	0.759

Table 2. Raw Agreement Rates of Consumer Car Trade-in Data Between Different Years

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
1999		0.914	0.899	0.900	0.872	0.861	0.858	0.855	0.847	0.847
2000			0.934	0.920	0.898	0.886	0.881	0.890	0.876	0.875
2001				0.934	0.914	0.896	0.894	0.894	0.882	0.881
2002					0.935	0.910	0.913	0.908	0.898	0.895
2003						0.944	0.936	0.927	0.909	0.914
2004							0.954	0.946	0.928	0.929
2005								0.954	0.931	0.932
2006									0.955	0.944
2007										0.954
2008										

prompt. We investigate this question further in the next section that explores prompt design.

3.3. Prompt Design

How should we structure the prompt to make the LLM generate better responses? To provide insight, we focus on the direct pairwise rating case where both the LLM and humans are asked to give a numerical similarity score between each pair of brands. We compare four different prompt structures shown in Online Appendix F.[9]

In simple prompts, we ask the LLM to generate a numerical similarity rating between each pair of brands; in few-shot prompts, we provide three examples (prompt-response pairs) selected from real human surveys and use a structure similar to the examples when asking for a numerical similarity rating; in *Role, Task, Format* (RTF) prompts, we specify that the LLM's role is a person who is filling out a survey, the task is to answer a question, and the format is to answer with an integer from 0 to 10; lastly, we try prompts where we add the RTF structure to the few-shot prompts. The exact wording of these prompts are documented in supplemental material F. As shown in Table 3, we find that when compared with human direct rating survey results, few-shot and RTF prompts lead to higher raw and adjusted agreement rates. We find that the combined prompt, which is both the longest and the most specific prompt, does not lead to responses that agree most often with human data.

We can draw two main insights from the previous analysis. From a managerial perspective, if the company's goal is to do some preliminary perceptual analysis without collecting any human data, the company can use RTF prompts to obtain quality LLM outputs that resemble real human responses. If the company already has a small set of human responses, the company can leverage these responses to potentially conduct a larger and broader study even with brands not included in the human survey using few-shot prompts.

Next, we expand our prompt analysis to compare LLM data with consumer car trade-in data. We calculate agreement rates between LLM-generated data and consumer trade-in data. Because car trade-in data are summarized quarterly totals, bootstrapping this data set is not possible, so we provide raw agreement rates in Table 4. Notably, few-shot and RTF prompts exhibit the highest agreement rates, mirroring the trend seen in the human survey data comparison.

The last column of Table 4 shows results where we tailor the RTF prompts to directly ask about consumer's car trade-in behavior: "Please imagine the following scenario. Act as if you own a car and want to trade it in for a similar car. If you currently have a [brand 1], how likely would you be to trade it in for an [brand 2]? Reply using only an integer from 0 (not at all likely) to 10 (very

Table 3. Agreement Rates of Different Prompts vs. Human Direct Rating Data

Type of prompts	Self-consistency rate	Agreement rate	Adjusted agreement rate
Simple	0.980 (0.971,0.988)	0.800 (0.795,0.804)	0.848 (0.843,0.853)
Few-shot	0.986 (0.971,0.988)	0.821 (0.820,0.830)	0.868 (0.867,0.877)
RTF	0.964 (0.947,0.978)	0.815 (0.810,0.823)	0.872 (0.866,0.880)
Combined	0.977 (0.967,0.989)	0.800 (0.794,0.808)	0.850 (0.844,0.859)

Notes. We collect 20 samples using each type of prompts. The self-consistency rates are the mean self-consistency rate of 200 equal-sized bootstrapped samples (20). The bootstrapped two-sided 95% confidence intervals are reported below the point estimates. Agreement rates and adjusted agreement rates are computed by fixing the human direct rating data and comparing each of the 200 bootstrapped samples against it. The reported point esimates of agreement rates and adjusted agreement rates are of the original samples (without bootstrapping).

Table 4. Raw Agreement Rates Between LLM-Generated Data Using Different Prompts and Consumer Car Trade-in Data

Year	Simple	RTF	Fewshot	Combine	Trade-in
1999	0.717	0.725	0.732	0.728	0.707
2000	0.718	0.738	0.742	0.735	0.727
2001	0.717	0.736	0.742	0.738	0.730
2002	0.718	0.739	0.735	0.734	0.726
2003	0.721	0.743	0.751	0.740	0.727
2004	0.733	0.756	0.763	0.753	0.730
2005	0.733	0.754	0.759	0.751	0.731
2006	0.741	0.751	0.771	0.759	0.735
2007	0.747	0.754	0.775	0.762	0.740
2008	0.740	0.759	0.776	0.763	0.728

likely). Please don't use any words in your response." Using a paired *t* test, we find that prompts that explicitly mention the trade-in scenario lead to responses that have significantly lower agreement rates than prompts that ask for overall similarity between brands.

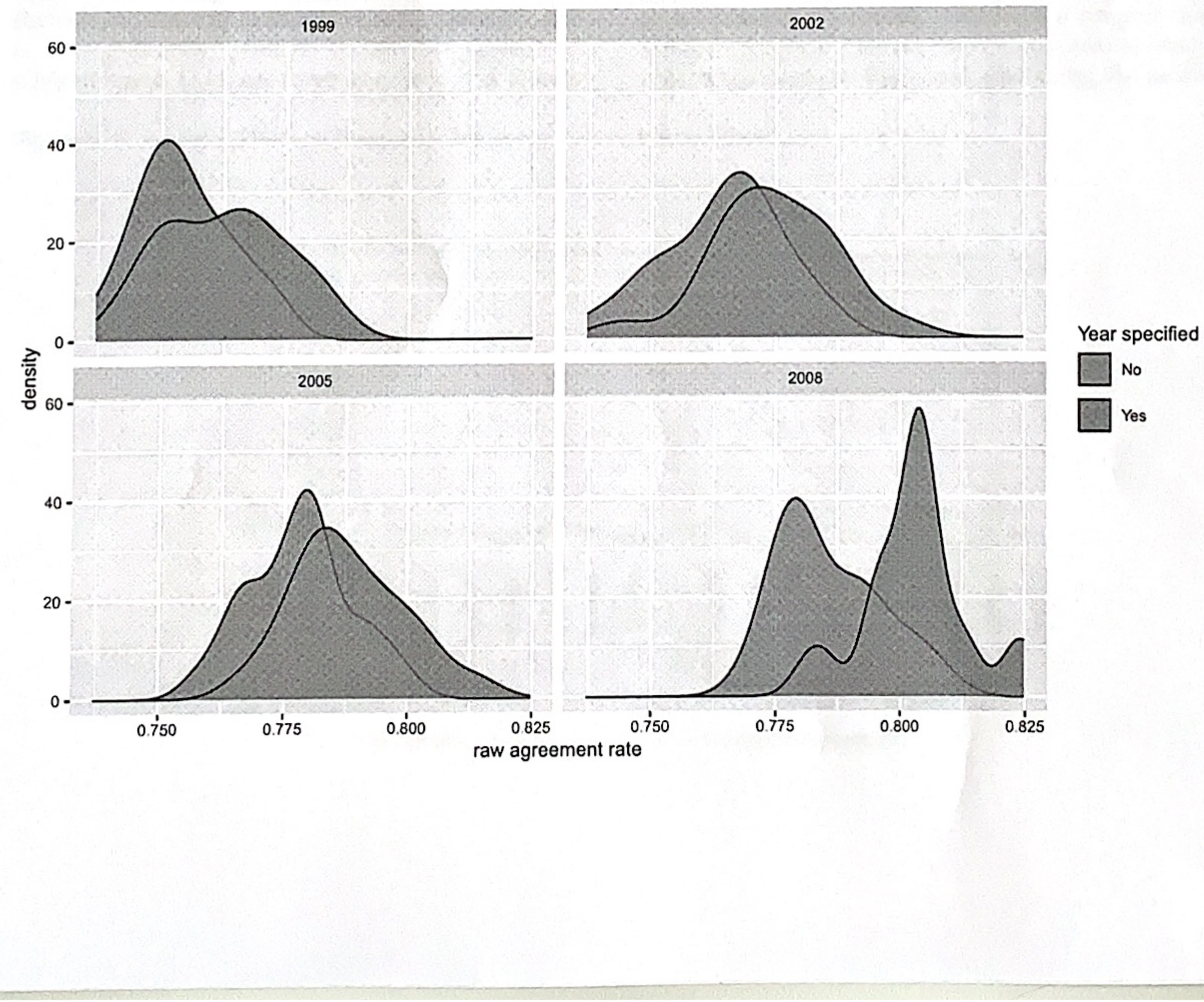
Overall, it appears RTF prompts are preferred due to high agreement with human data and not needing human input. Interestingly, specifically mentioning the "trade-in" does not result in a better match with trade-in data, perhaps due to the more context-specific use of the word.

Finally, we analyze how adding a year specification to the LLM pairwise rating prompts affects agreement rates between LLM and consumer trade-in data. We modify the RTF prompts to "I want you to act as a person filling out a survey in [*year*]. I will ask you a question and you must answer using only an integer, no words. You will reply with an integer between 0 and 10. My first question is in your opinion, how similar are the car brands a and b on a scale of 0 to 10 where 10 means very similar?" We gather LLM survey data for specific years (1999, 2002, 2005, 2008) and calculate agreement rates with consumers' trade-in data for each year. We compare these rates to those of non-year-specific LLM data. Figure 4 shows the distribution of agreement rates between the LLM data and the trade-in data for the cases when the year is specified in the prompt and when it is not. For all four years under examination, the year-specific LLM appears to aligning more closely with trade-in data. We calculate the increase in raw agreement to be a significant 1.14%. This analysis suggests that specifying the year in the prompt allows us to extract better information from the LLM that is more representative of perceptions in that time period.

3.4. Problematic Brands

An important and practical extension is to identify the top problematic brands: brands for which the LLM responses do not reflect the human survey data.

Figure 4. (Color online) Distribution of Sample Raw Agreement Rates Between Year-Specific Consumer Trade-in Data and Year-Specific LLM Data vs. LLM Data Without Year Specification



Researchers and practitioners can choose to omit these brands when extracting insights using LLM or do further human studies on the subset of problematic brands.

We identify problematic brands by leveraging a key insight: When LLMs and humans differ in their perception of a brand, it suggests that LLM's self-consistency regarding that brand is likely lower. Removing such brands from the data set should enhance the overall self-consistency of LLM data. This guides us in creating a metric for brand-level self-consistency. Specifically, we generate 20 pairs of equally sized bootstrapped samples from the GPT4 rating data set. For each pair, we exclude brand i and calculate the raw agreement rate between the two sets (the self-consistency rate without brand i). The average of these 20 rates represents the brand-level self-consistency measure. If the without-brand self-consistency rate is significantly higher than the overall self-consistency rate, brand i has a detrimental impact on the overall self-consistency, indicating the LLM's inconsistency in perceiving brand i. Similarly, we can pinpoint brands responsible for major LLM-human disagreements by computing the without-brand agreement rate for each brand.

Figure 5 plots the without-brand LLM self-consistency rate on the top and the without-brand LLM human agreement rate in the bottom. The without-brand self-consistency rates are plotted in decreasing order, and we can observe that the two most problematic brands (brands with the highest without-brand agreement rates) Renault and Subaru are identified, and statistically, there is a significantly positive correlation between the without-brand LLM self-consistency rate and without-brand LLM human agreement rate.[10] Using this study, we show that we can identify the most problematic brands in a LLM-powered study by calculating the without-brand LLM self-consistency rate of each brand. This solely LLM-driven investigation tool enables marketers to make more detailed decisions when selecting brands to include in further studies.

4. Introducing Consumer Heterogeneity

This section explores the impact of consumer heterogeneity on LLM-generated consumer research. Our objective is twofold. First, we want to provide further external validity to our empirical studies on perceptions. As mentioned earlier, as opposed to preferences, brand perceptions are *not* supposed to be different across consumer segments. A sports car is seen as "sporty" by all consumers, not just those who purchase it. We confirm this hypothesis empirically for both LLM-generated and human data sets: Consumer heterogeneity does not have a significant effect on generating perceptual maps (for more detail, please refer to the heterogeneity section in Online Appendix E). Second, we want to show that, LLM-based consumer research has the capacity to meaningfully explore consumer heterogeneity. We do so by turning to consumer *preferences*, which we expect to exhibit significant consumer heterogeneity.

We collect human survey data asking for participants' favorite car brands. The exact prompt is "What is your favorite car brand?," and participants are allowed to give open-ended answers.[11] We have a roughly even number of male and female respondents, and we evenly split respondents to (high and low) along the income

Figure 5. (Color online) Without-Brand Self-Consistency Rate vs. Without-Brand Agreement Rate

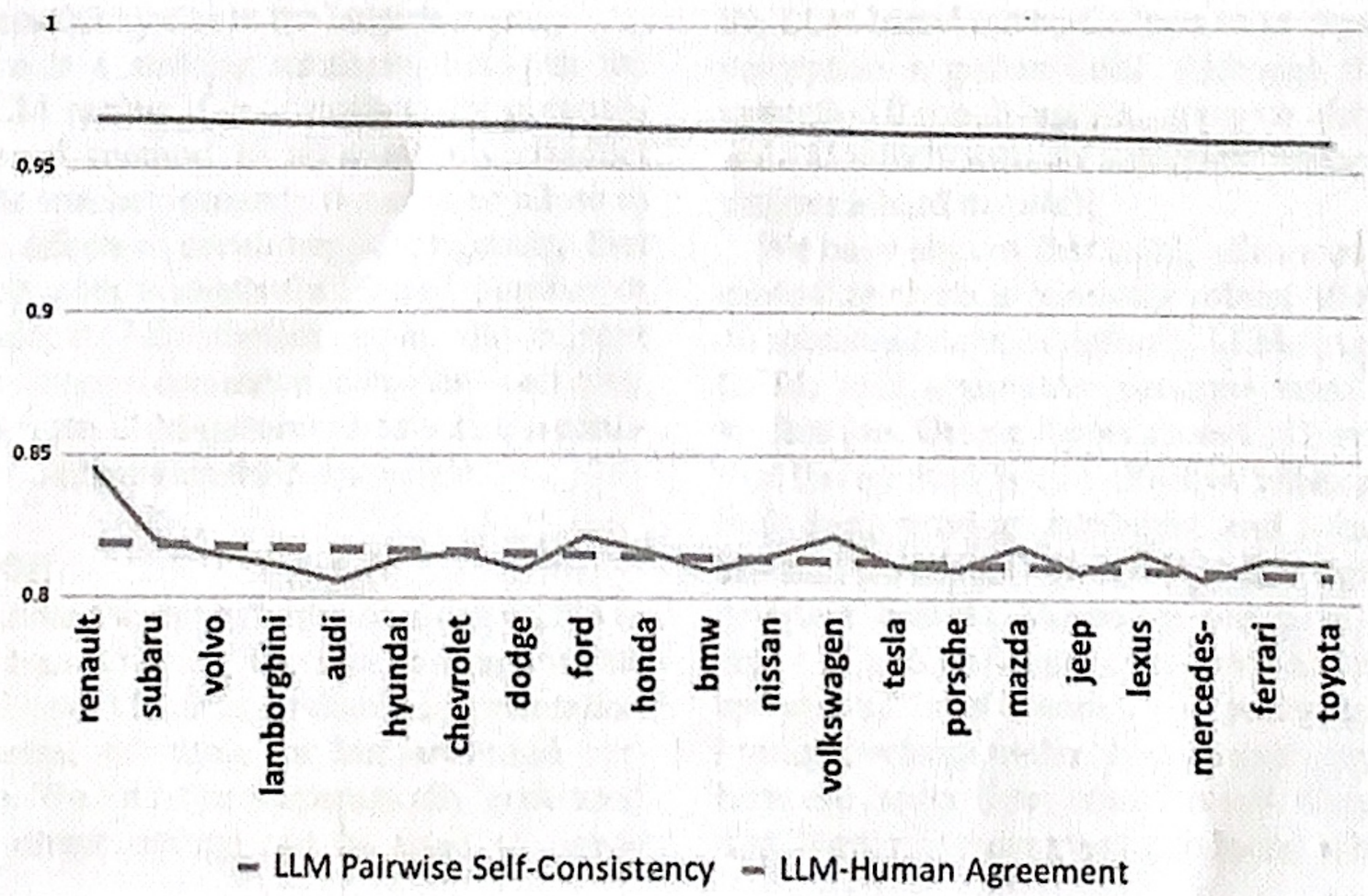
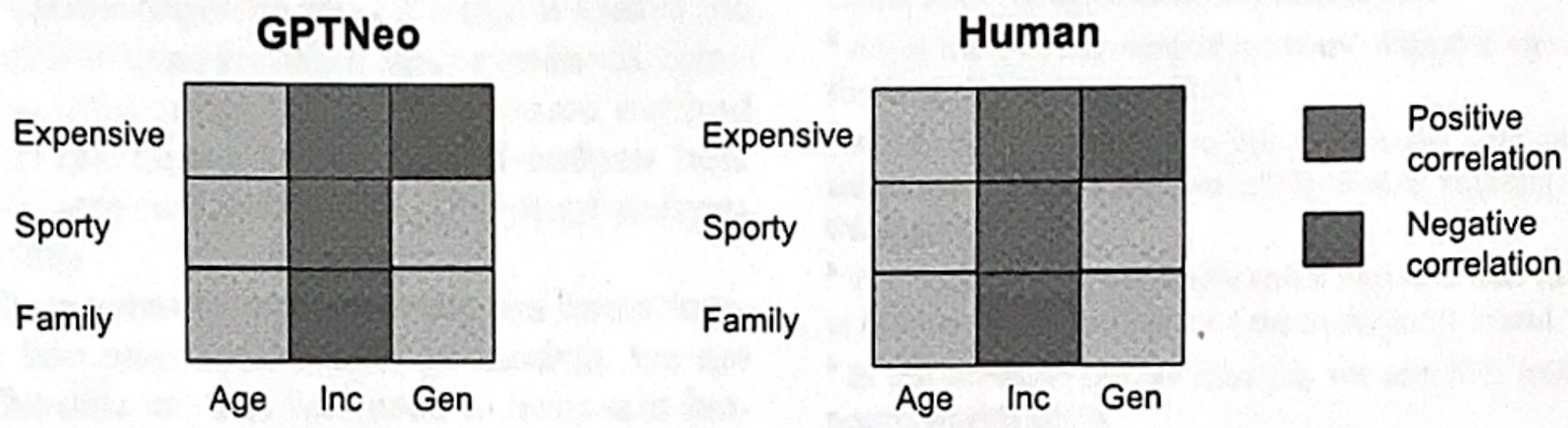


Figure 6. (Color online) Consumer Characteristics That Significantly Correlate with Preferences in Both Data Sets



Note. Gray squares represent insignificant correlations in at least one of the data sets; Inc, Income; Gen, Gender; see exact coefficients in Online Appendix E.

and the age dimension. This balanced data set eliminates the concern for multicollinearity between age and income. The three consumer variables result in $2^3 = 8$ subsets. Overall, we have 658 human participants with at least 70 in each bucket.

For the LLM data, we use prompts like "A young and poor male's favorite car brand is ..." or "A young and poor female's favorite car brand is ...," and so on. We collect 6,000 LLM responses for each category of consumers. Then we run linear regressions to test whether age, income, or gender have significant correlations with consumers' preference for expensive cars, sports cars, and family cars.

Figure 6 shows the comparison between LLM and human preference results. The results are strikingly similar: Every coefficient that is significant for both data sets is directionally the same. In the human data, income and gender significantly affect brand preference. For example, wealthy consumers prefer sports cars and expensive cars, whereas low-income consumers prefer family cars. We recover all but one of the statistically significant human effects using the LLM data. In addition, The LLM data yield a higher number of significant results (e.g., for age), most likely due to the larger sample.

Overall, there is a striking similarity between the human and LLM results demonstrating the potential of the LLM-based method to economically conduct segment-specific market research. It may also allow to find qualitative effects of consumer heterogeneity that are not detected with a relatively limited number of human respondents. Nonetheless, given the current technology, we do not recommend using the exact coefficients derived from LLM-generated data as the confidence intervals do not match the human data.

5. Conclusion

Our paper explores the opportunity of using LLMs to conduct perceptual analysis. The results demonstrate that using LLM-based tools is a reliable augmentation or even a potential substitute for human brand perception surveys. We find that automatically generated sentences and direct ratings can be used to create perceptual maps that broadly match those created from human surveys. Nonetheless, the most powerful applications of such methods are likely to be in conjunction with human surveys, where an initial analysis is first conducted with LLMs. That, in turn can inspire the design of proper human surveys, which can validate the results. The basic steps of our method are the following: (1) choose a LLM whose learning corpus is suitable for the application or fine-tune a model with relevant data; (2) choose a broad set of brands in the category and collect responses using RTF prompts; (3) determine the problematic brands using without-brand self-consistency tests; (4) conduct perceptual analysis on nonproblematic brands; and (5) if needed, collect human data to validate the results and focus on the problematic brands.

One important opportunity we identify is using LLMs to answer more detailed questions about brands. We find that we can replicate survey differences along demographic variables collected in human surveys and survey differences across time. The more refined the definition of the variables we use to split consumers into different cells, the more difficult it is to collect human response. Hence, the LLM-based method allows to understand consumer perception in greater detail. Although the possibility of validation through human surveys is always present, the scale at which such an automatic method can generate insights is hard to match.

We have shown that using advanced LLMs for perceptual analysis is relatively robust. We have reported on robustness across different LLMs (e.g., GPTNeo and GPT4) and alternative prompts used. In additional studies (see Online Appendices C, D, and G), we show that the method is also effective when perceptions are built from product attributes, and it can replicate the abstract perceptual dimensions that consumers use to compare brands. Moreover, we have replicated our study in additional product contexts (namely, apparel brands and hotel brands). We found that, generally car brands perform better than apparel brands, which in turn are better than hotel brands. Our hypotheses is that higher involvement products, with more stable

preferences are better suited for our methods than their lower involvement counterparts, although we leave the verification of this hypothesis for future research. Similarly, we have only compared our LLM-based method to human surveys. Future research could evaluate how such methods fare compared with perceptual analysis using UGC data.

Despite all its benefits, our approach has limitations. Because we use pretrained language models, we are limited by the data set that was used to train said language models (Atari et al. 2023). The trend is for LLMs to use increasingly large and broad training data inclusive of information on all aspects of life including market information. Nonetheless one cannot guarantee the presence of relevant data in the model unless custom data are specifically included in the training set or the LLM is augmented with a retrieval capability. Possibly the biggest limitation of current language models is the time dimension. Today, the state-of-the-art language models use their entire training corpus to generate continuations instead of explicitly focusing on texts from a certain time period. The model itself can learn to interpret reference to time in the text (as we show in our temporal analysis), but thus far, performance on this front has not been a priority. Although the inability to understand how perceptions change over time is a serious limitation, there is no theoretical obstacle in the way of training language models that take time into account. We expect significant progress on this front in the coming years. We believe that future work on the use of large LLMs in the context of marketing is warranted.

Acknowledgments

The authors thank Florian Stahl for providing the car trade-in data for our analysis; participants at the University of Washington Marketing Camp, the Wharton Business & Generative AI Workshop, and the MSI 2023 immersion event for comments; and the anonymous reviewers who helped improve the paper. The last three authors are listed in alphabetical order.

Endnotes

[1] As described later, we explore alternative prompts to gauge the robustness of the responses. Indeed, as in other contexts, so-called, "prompt engineering" is an important aspect of using these language models at the current stage, although ongoing and future research reduce this need.

[2] The specific models we use are the fully open source GPTNeo project developed by Black et al. (2021) and the state-of-the-art ChatGPT (version based on GPT4) released by OpenAI, but our methods are flexible to the substitution of any other large (generative) language model.

[3] One of the main benefits of using an ordinal embedding method is that, as shown by Kleindessner and Luxburg (2014), there exists a unique embedding for each data set under our assumptions.

[4] For further technical details, see Online Appendix B.

[5] In the rare instances of triplets where $F(i,j) = F(i,k)$, we drop the triplet from the agreement rate calculations.

[6] All of the exact agreement rates and adjusted agreement rates are shown in Online Appendix C.

[7] We do not have access to portions of this data set that were collected after 2008; hence, we are limited to running the analysis for this time period.

[8] We conduct a simple sanity check and find that more than 50% of consumers trade in a car for a car in the same brand.

[9] In the previous studies thus far, we use RTF (role, task, format) prompts with GPT4.

[10] The correlation is 0.37, and the one-sided p value is 0.048.

[11] Of all participants, only seven indicated that they did not have a favorite car brand, and only one indicated more than one brand.

References

Atari M, Xue MJ, Park PS, Blasi DE, Henrich J (2023) Which humans? Working paper, Harvard University, Cambridge, MA.

Berger J, Packard G, Boghrati R, Hsu M, Humphreys A, Luangrath A, Moore S, et al. (2022) Marketing insights from text analysis. *Marketing Lett.* 33(3):365–377.

Black S, Gao L, Wang P, Leahy C, Biderman S (2021) GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. https://doi.org/10.5281/zenodo.5297715.

Brand J, Israeli A, Ngwe D (2023) Using GPT for market research. Working paper, Harvard Business School, Cambridge, MA.

Culotta A, Cutler J (2016) Mining brand perceptions from twitter social networks. *Marketing Sci.* 35(3):343–362.

Dzyabura D, Peres R (2021) Visual elicitation of brand perception. *J. Marketing* 85(4):44–66.

Espejel JL, Ettifouri EH, Mahaman SYA, Chouham EM, Dahhane W (2023) Gpt-3.5, gpt-4, or bard? Evaluating llms reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing J.* 5:100032.

Horton JJ (2023) Large language models as simulated economic agents: What can we learn from homo silicus? NBER Working Paper No. 31122, National Bureau of Economic Research, Cambridge, MA.

Humphreys A, Wang RJ-H (2018) Automated text analysis for consumer research. *J. Consumer Res.* 44(6):1274–1306.

Jain L, Jamieson K, Nowak R (2016) Finite sample prediction and recovery bounds for ordinal embedding. Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, eds. *Proc. NeurIPS* (Curran Associates, Inc., Red Hook, NY), 2720–2728.

Kleindessner M, Luxburg U (2014) Uniqueness of ordinal embedding. Balcan MF, Feldman V, Szepesvári C, eds. *Proc. 27th Conf. on Learn. Theory*, vol. 35 (PMLR, New York), 40–67.

Liu J, Toubia O (2018) A semantic approach for estimating consumer content preferences from online search queries. *Marketing Sci.* 37(6):930–952.

Liu L, Dzyabura D, Mizik N (2020) Visual listening in: Extracting brand image portrayed on social media. *Marketing Sci.* 39(4): 669–686.

Matthe M, Ringel DM, Skiera B (2022) Mapping market structure evolution. *Marketing Sci.* 42(3):589–613.

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

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, et al. (2022) Training language models to follow instructions with human feedback. Koyejo S, Mohamed S, Agarwal A,

Belgrave D, Cho K, Oh A, eds. *Proc. NeurIPS* (Curran Associates, Inc., Red Hook, NY), 1–15.

Radford A, Narasimhan K (2018) Improving language understanding by generative pre-training. Working paper, OpenAI, San Francisco, CA.

Timoshenko A, Hauser JR (2019) Identifying customer needs from user-generated content. *Marketing Sci.* 38(1):1–20.

Tirunillai S, Tellis GJ (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *J. Marketing Res.* 51(4):463–479.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, et al. (2017) Attention is all you need. Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Proc. NeurIPS* (Curran Associates, Inc., Red Hook, NY), 6000–6010.