Supplemental Materials

We present the supplemental materials in seven sections. In section A, we define all of the technical terms and notations used in this paper. This is a useful reference for understanding the mathematical details of this paper. In section B, we discuss the mathematical formulation of our methods and some properties of these methods. In section C, we present supplemental materials for the car brand study. In section D, we document the full apparel study. In section E, we add some details of the heterogeneity study. In section F, we show all of the prompts we tested for the prompt selection study using car brands. In section G, we show a study of the applicability of the LLM-powered approach to conduct perceptual analysis in different product categories. In section H, we test whether the order of brands carry any information about relative similarities in the case where the open-ended LLM responses contain multiple brands. In section I, we comment on two expository points about LLM generated data: what is likely making LLMs sensitive to a time specification in the prompt and why the self-consistency rates of LLM generated open-ended data is high. In section J, we show perceptual maps draw from LLM pairwise data with year specification and car trade-in data by year.

A Definitions and notations

Term	Definition
LLM	Two large language models GPTNeo and ChatGPT-4
Open-ended prompts	Prompts that ask human participants or LLMs to generate a continuation after each prompt.
Open-ended responses	Responses to open-ended prompts
Pairwise numerical rating prompts	Prompts that ask human participants or LLMs to generate a numerical answer to each prompt. The pairs may be a pair of brands or a brand and an attribute
Pairwise numerical rating responses	Responses to pairwise numerical rating prompts.
Frequency matrix	A matrix whose i, jth entry is the number of times responses containing brand j appear after the ith prompt.
Baseline value	An implicit value of each brand that represents its top-of-mind awareness.
Baseline effect	A bias in the frequency table that is caused by brands baseline values.
Triplet	A triple of elements in a set. The triplet is invariant to a reordering of the last two elements i.e. $(A, B, C) \equiv (A, C, B)$
Triplet evaluation function	A function that takes a triplet as an input and outputs 1 if the first two elements are more similar than the first and third elements, and outputs 0 if the first two are less similar.
Agreement rate	For any two data sets with the same set of triplets, the agreement rate is the probability that take any triplet, the evaluation outputs based on both data sets are the same
Self-consistency rate	The agreement rate between two bootstrapped samples of the same data set.

Table S1: List of definitions

Term	**Notation**
B	The set of brands in a study. B_{-ij} denotes the set of brands except i and j.
$H(i,j)$	The hidden true similarity between a pair of brands $i, j \in B$
b_i	The baseline value of brand $i \in B$
$s_D(i,j)$	The imputed similarity score between brands $i, j \in B$ based on data set D.
D	Any data set used in the study like LLM text completion, human direct rating etc.
F_D^B	The observed similarity measure of brands in a data set D. For any $i, j \in B$, $F_D^B(i,j) = G_D(H(i,j), b_i)$
eval	The triplet evaluation function. For $i, j, k \in B$, $$(\mathbf{i}, j, k) = \begin{cases} 1, & F(i,j) > F(i,k) \\ 0, & F(i,j) < F(i,k) \end{cases}$$
AR$(\cdot,\cdot)$	The agreement rate between two data sets.
SAR$(\cdot)$	The self-consistency rate of a data set.

Table S2: List of notations

B Methods

Our methods revolve around calculating similarity scores of brands in the same data set and calculating the similarity score of different data sets. To motivate our method to calculate brand similarity scores in the same data set, we first describe one of the main challenges in imputing brand similarity scores using text data–the baseline effect. As defined in section A, difference in baseline values appear when the set of brands have different levels of popularity. This causes ambiguity when we directly use frequency counts of brand names as a measure of similarity. A well-known brand like Ford may appear frequently following all of the prompts, but this is not an indication that Ford is similar to all car brands, rather, this is in part because Ford in highly linked to the car category; therefore, when the LLM or human are prompted to answer an open-ended prompt with the name of a car brand, the brand Ford comes to their minds automatically. This baseline effect is not significant in direct rating responses; therefore, our method focuses on the case where we have open-ended text responses.

Therefore, we propose the following method to calculate the similarity score between brands in the same open-ended data set. For any set of brands B and a set of open-ended responses D, we count how many times each brand appears after each prompt. We use this as the observed brand similarity measures F_D^B. More specifically, if "Audi" appears 10 times after the prompt "The car brand BMW is similar to the car brand" in a data set D, then $F_D^{\text{car brands}}(\text{Audi}, \text{BMW}) = 10$. For any brands $i, j \in B$, the similarity score of brands i and j based on D is computed as

$$s_D(i,j) = \text{mean}_{k,l \in B_{-ij}, k \neq l}[\text{eval}(i,k,l) == \text{eval}(j,k,l)],$$

where the evaluation function is specified in table S2.

Moreover, if we assume there is a true hidden similarity $H(i,j)$ between each pair of brands i and j, and the observed frequency $F_D^B(i,j)$ and $F_D^B(j,i)$ are monotonically increasing with respect to the hidden true similarity of the two brands, we can show that the evaluation functions are independent of the baseline values of brands i and j. Therefore, when we take the mean of these evaluation functions, the imputed similarity score $s_D(i,j)$ is also independent of these baseline values.

Property 1. Under the monotonicity assumption, for any triplet $t = (i, j, k)$, the value of $\text{eval}(t)$ does not depend on b_i. Mathematically, this property can be written as

$$\text{eval}(t) = 1 \Leftrightarrow \text{H}(i, j) > \text{H}(i, k).$$

Proof (property 1): First if $\text{H}(i, j) = \text{H}(i, k)$, we have $F_D^B(i, j) = F_D^B(i, k)$ and t will be dropped. Thus, we only have to consider the two possibilities $\text{H}(i, j) > \text{H}(i, k)$ and $\text{H}(i, j) < \text{H}(i, k)$ in this proof.

Assume $\text{H}(i, j) + \delta = \text{H}(i, k) \in \mathbb{R}$ for some nonnegative δ. Take any $b_i \in \mathbb{R}$, under the monotonicity assumption, let $F_D^B(i, j) = G_D(\text{H}(i, j), b_i)$ which monotonically increases with respect to $\text{H}(i, j)$ the hidden true similarity between i, j

$$F_D^B(i, j) = G_D(\text{H}(i, j), b_i) = G_D(\text{H}(i, k) + \delta, b_i)$$

$$\geq G_D(\text{H}(i, k), b_i) = F_D^B(i, k)$$

$$\implies \text{eval}(t) = 1$$

Thus, we have $\text{H}(i, j) > \text{H}(i, k) \implies \text{eval}(t) = 1$.

Now, we prove the converse,

$$\text{eval}(t) = 1 \implies F_D^B(i, j) > F(i, k)$$

$$\implies G_D(s_D(i, j), b_i) > G_D(s_D(i, j), b_i)$$

$$\implies \text{H}(i, j) > \text{H}(i, k)$$

Therefore, $\text{eval}(t) = 1 \implies \text{H}(i, j) > \text{H}(i, k)$. Similarly, we can show the other direction $\text{H}(i, j) < \text{H}(i, k)$ □

To see this method in action, we present a toy example to walk through the calculations. We use a simple example with just 4 car brands: Ford, Chevrolet, Mercedes, and BMW. We present a hypothetical frequency table of these four brands and show how ordinal embedding based similarity

scores can be computed.Then we show another hypothetical frequency table that represents the results from another data set. We compare these two frequency tables to show how the triplet method can be applied to calculate data set level similarity scores.

Table S3: Hypothetical frequency table representing LLM data

	Ford	Chevrolet	Mercedes-Benz	BMW
Ford	–	15000	7600	9400
Chevrolet	490	–	24	73
Mercedes-Benz	690	100	–	2800
BMW	1400	110	3500	–

Table S3 shows the hypothetical results representing the LLM data. Columns correspond to prompts and rows correspond to frequency counts. For example, the number 15000 means that for the prompt containing Chevrolet, we counted 15000 mentions of Ford. Glancing through the numbers, it appears Ford and Chevrolet are similar to each other, and Mercedes and BMW are similar to each other whereas relatively speaking the two pairs are less similar. But one must take into account that Ford has the largest baseline value (approximated by the row sum), Mercedes and BMW have medium baseline values, and Chevrolet has the smallest baseline value. Henceforth, given a prompt containing Mercedes, even though Ford is mentioned more times (7600) than BMW (3500), we should infer a higher similarity between Mercedes and BMW than between Mercedes and Ford.

Our similarity score controls for the baseline. To calculate the similarity score for Ford and Chevrolet, we first compute the evaluation function for Ford and Chevrolet respectively. Here, since we only have 4 brands, there is only one evaluation function for both Ford and Chevrolet, and the evaluation function operates on (X, BMW, Mercedes-Benz) where X is Ford and Chevrolet, respectively.

We have

$$\text{eval}(\text{Ford}, \text{BMW}, \text{Mercedes-Benz}) = 1$$

and

$$\text{eval}(\text{Chevrolet}, \text{BMW}, \text{Mercedes-Benz}) = 1.$$

Therefore,

$$s(\text{Ford}, \text{Chevrolet}) = [\text{eval}(\text{Ford}, \text{BMW}, \text{Mercedes-Benz}) == \text{eval}(\text{Chevrolet}, \text{BMW}, \text{Mercedes-Benz})]$$

$$= \mathbb{1}(1 == 1) = 1.$$

Similarly, we can calculate the similarity score between Ford and BMW:

$$\text{eval}(\text{Ford}, \text{Chevrolet}, \text{Mercedes-Benz}) = 1$$

and

$$\text{eval}(\text{BMW}, \text{Chevrolet}, \text{Mercedes-Benz}) = 0.$$

Therefore,

$$s(\text{Ford}, \text{BMW}) = [\text{eval}(\text{Ford}, \text{Chevrolet}, \text{Mercedes-Benz}) == \text{eval}(\text{BMW}, \text{Chevrolet}, \text{Mercedes-Benz})]$$

$$= \mathbb{1}(1 == 0) = 0.$$

We can use the same steps to get $s(\text{BMW}, \text{Mercedes-Benz}) = 1$, $s(\text{Ford}, \text{Mercedes-Benz}) = 0$, $s(\text{Chevrolet}, \text{BMW}) = 0$, and $s(\text{Chevrolet}, \text{Mercedes-Benz}) = 0$. Therefore, we can recover that BMW and Mercedes are similar, and they are less similar to Ford and Chevrolet. Ford and Chevrolet are similar, and they are less similar to BMW and Mercedes.

The second major component of our methodological contribution is the nonparametric triplet method we develop to calculate similarity scores of data sets. Formally, for any two data sets D_1 and D_2 that contain the set of brands B, the agreement rate between the two data sets is

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

In simpler terms, the agreement rate of $F^B_{D_1}$ and $F^B_{D_2}$ is the probability that both data sets agree on the position of a brand relative to two other brands. The self-consistency rate of F^B_D is formally expressed as

$$\text{SAR}(F^B_D) = \text{AR}(F^B_{D_x}, F^B_{D_y})$$

where $F^B_{D_x}$ and $F^B_{D_y}$ are the observed similarity measures of two equal-sized resamplings of D. We show that the expectation of the agreement rate between two data sets is at most the average of the expectation of the self-consistency rate of the two data sets.

Property 2.

$$\mathbb{E}\left[\text{AR}(F^B_{D_1}, F^B_{D_2})\right] \leq \frac{\mathbb{E}\left[\text{SAR}(F^B_{D_1})\right] + \mathbb{E}\left[\text{SAR}(F^B_{D_2})\right]}{2}.$$

Proof (property 2): Let $p_t^{D_1}$ be the probability that for a given triplet t, the function eval$(t) = 1$ for any resampling of $F^B_{D_1}$. We define $p_t^{D_2}$ similarly for the distribution $F^B_{D_2}$. We can then express

$$\mathbb{E}\,\text{SAR}(F^B_{D_1}) = \frac{1}{|T|}\sum_t (p_t^{D_1})^2 + (1 - p_t^{D_1})^2,$$

and a corresponding expression holds for $\mathbb{E}\,\text{SAR}(F^B_{D_2})$. Assuming resamplings of $F^B_{D_1}$ and $F^B_{D_2}$ are taken independently, the resampled agreement rate between the two data sets can be written as

$$\mathbb{E}\,\text{AR}(F^B_{D_1}, F^B_{D_2}) = \frac{1}{|T|}\sum_t (p_t^{D_1})(p_t^{D_2}) + (1 - p_t^{D_1})(1 - p_t^{D_2}),$$

For any t, we have

$$2(p_t^{D_1})(p_t^{D_2}) + 2(1 - p_t^{D_1})(1 - p_t^{D_2}) \leq (p_t^{D_1})^2 + (p_t^{D_2})^2 + (1 - p_t^{D_1})^2 + (1 - p_t^{D_2})^2,$$

completing the proof. □

C Car study – overall similarity and attribute-based study

We elaborate on three topics of the car study in the main paper. First, we provide a detailed description of the list of car brands that are studied and the set of participants in our human surveys. Then, we provide a table that shows the exact numbers in our overall similarity analysis of car brands. Lastly, we show details of an analysis that conducts perceptual analysis by comparing key attributes of car brands.

We study a set of 21 major car brands as shown in Table S4. This list contains brands from all major production regions: U.S., Europe, and Asia, and this list also contains a large variety of car brands: sports car brands, daily commuter car brands, and luxury non-sports brands. This list is selected by ranking the brands with the highest frequency when the GPTNeo is asked to respond to the prompt "The car brand...". Overall, these car brands contributed to over three quarters of the total number of cars sold in the United States in 2021.

Table S4: Table of car brands

Brand	Region	Number of vehicles sold (2021)
Toyota	Asia	1,933,099
Ford	United States	1,804,793
Chevrolet	United States	1,468,889
Honda	Asia	1,308,476
Nissan	Asia	919,090
Jeep	United States	768,713
Hyundai	Asia	726,715
Subaru	Asia	598,480
Volkswagen	Europe	366,462
BMW	Europe	336,694
Mazda	Asia	332,821
Mercedes-Benz	Europe	329,665
Lexus	Asia	304,476
Tesla	United States	301,998
Dodge	United States	215,726
Audi	Europe	196,038
Volvo	Europe	122,173
Porsche	Europe	70,025
Lamborghini	Europe	2,942
Ferrari	Europe	2,831
Renault	Europe	NA[1]

In our study, we collect three sets of human responses: open-ended human responses to prompts like "the car brand X is most similar to the car brand...", numerical direct rating prompts like "how similar are the car brands X and Y?", and attribute-based direct rating prompts like "on a scale of 0 to 10, how sporty is the car brand X?". The information of participants in these three surveys are listed in tables S5,S6, and S7

Table S5: Human open-ended data descriptive statistics (car overall similarity)

	Education	Income	Response duration (sec)	Language
Mean	4.49	6.94	198.24	0.995
Standard Deviation	1.43	3.51	177.63	–
Median	5	7	137	1
1st decile	2	2	72.0	0
9th decile	6	12	397.9	1
Min	1	1	10	0
Max	8	12	1288	1

The language indicator is 1 if the language is English, and it is 0 for all other languages.

Table S6: Human direct rating data descriptive statistics (car overall similarity)

	Education	Income	Response duration (sec)	Age	Gender
Mean	4.26	6.49	104.90	31.09	0.525
Standard Deviation	1.44	3.55	245.42	10.97	–
Median	5	6	76	28	1
1st decile	2	2	50.9	20	0
9th decile	6	11	144.1	46	1
Min	1	1	32	18	0
Max	8	12	5273	78	3

In the gender column, 0 stands for male, 1 stands for female, 2 stands for non-binary, and 3 stands for rather not say.

[1]Renault is not in the U.S. market but has a large global presence.

Table S7: Human open-ended data descriptive statistics (car attribute-based study)

	Gender	Age	Response duration (sec)	Language
Mean	0.89	25.22	54.27	1
Standard Deviation	–	7.00	59.24	–
Median	1	23	45	1
1st decile	0	20	28	1
9th decile	1	32	84	1
Min	0	19	17	1
Max	3	77	1703	1

The language indicator is 1 if the language is English, and it is 0 for all other languages. In the gender column, 0 stands for male, 1 stands for female, 2 stands for non-binary, and 3 stands for rather not say.

As shown in Table S5, out of the 402 participants in our open-ended car survey, the mean and median are both within 4 to 5, which corresponds to associated and bachelor degrees. In addition, we observe the middle 80% of participants ranges from education level 2 to level 6, which corresponds to high school graduates and master degree holders. The mean and median of income are both around level 7, which corresponds to an annual income of 60 to 70 thousand which agrees with the average U.S. annual household income of 67.5 thousand dollars. Moreover, the distribution of household income is skewed to the right with the first decile at 10 to 20 thousand dollars per year, and the nineth decile is at above 150 thousand dollars per year. This means our data is able to capture some information about high income individuals too. In addition, we observe that the mean and median of response time are both 2 to 3 minutes, and the middle 80% of participants take 1 to 7 minutes to complete the survey. Lastly, we observe that over 99.5% of participants chose English as their preferred language, so there is no significant language barrier in our study.

Comparing the education level and annual household income in Table S6 and Table S5, the distribution of participants in the car direct similarity rating survey is roughly the same as the distribution of participants in the open-ended car survey. We have additional demographic information in the direct rating data. More specifically, we observe that the number of males and females is roughly equal in our study, and most of the participants are young to mid-aged adults. The age group represented by these samples is active in the car market. In addition, note that on average, it takes less time for participants to respond to direct rating questions than open-ended questions.

This agrees with our intuition.

Table S7 shows the descriptive statistics of participants in the attribute-based survey. The mean and median age of our participants are between 23 to 26 years old. The mean and median response time of our participants was almost 1 minute. In addition, over 80% of our participants are female. Note that we show in the consumer heterogeneity section, consumer's perceptions are do not change much across gender especially for younger participants. Furthermore, all of our participants chose English as their preferred language, indicating that there is no language barrier in the data collection process.

Overall, in all three surveys, we have participants that cover a wide range of demographics, and there is no language barrier in any of the surveys.

We see in the main body of this paper that all LLM human car brand perception adjusted agreement rates are above 75%. In tables S8, we show the exact numbers and the self-consistency rate of each data set in the overall similarity study.

Table S8: Triplet agreement rates of car brand data sets in the overall similarity study

	LLM open	Human open	LLM pairwise (GPT4)	Human pairwise
LLM open	.955 (.949,.962)	.801 .749	.774 .736	.773 .719
Human open	–	.916 (.893,.933)	.859 .807	.893 .813
LLM pairwise (GPT4)	–	–	.964 (.947,.978)	.872 .815
Human pairwise	–	–	–	.905 (.879,.935)

Black numbers show the point estimates of the self-consistency rates and agreement rates between data sets. Red numbers are agreement rates between data sets after adjusting for imperfect self-consistency. The numbers in the parentheses are bootstrapped 95% confidence intervals of self-consistency rates.

Another approach to construct perceptual maps is to start with key attributes that define brands in a market. For example, some important attributes in the car market are sportiness, fuel-efficiency etc[2]. We adopt our method and apply it to an attribute-based study of car brands.

[2]In this study, we include 12 attributes that we identified by running a prompt that allows the LLM to broadly

The first type of data are LLM-generated open-ended data (using GPTNeo-2.7B). We first construct a numerical attribute-brand frequency matrix by counting the number of times each brand appeared following each prompt containing the attribute. For example, if the brand BMW appeared a total of 10 times in all the completions we collected for the prompt "The most eco-friendly car brand is...", the entry corresponding to (BMW, eco-friendly) would be 10. Each row in the numerical matrix represents a brand, and each column represents an attribute. We collect 21000 responses using each prompt. We regard each row as a data point in a Principal Component Analysis (PCA). To avoid the possibility that all of the PCA results will be dominated by a few brands with high baseline values, we scale each data point to a unit vector. We can then extract meaningful features from the loadings of attributes.

In addition, we collect human- and LLM-generated pairwise numerical rating data for the attribute-based analysis. We do this by directly asking participants (and GPT-4) to rate brands on a set of attributes[3]. Similar to the LLM open-ended analysis, we organize the ratings of the same brand in the same row and scale each row to a unit vector. Then, we run PCA on this matrix to summarize attributes into features (factors). Once again, given these are direct questions at the brand level, we do not encounter a baseline problem with this type of human data.

With these three data matrices (LLM open-ended, LLM pairwise numerical rating, and human pairwise numerical rating), we can calculate data set level similarity for attribute-based analysis results using the triplet method to calculate a similarity score between LLM and human data. Different from the overall similarity study, each triplet includes a brand and two attributes, instead of three brands. Intuitively, the triplet score tells us that given a brand and two attributes, what is the probability that both human and LLM think any other brand is higher on attribute 1 or higher on attribute 2.

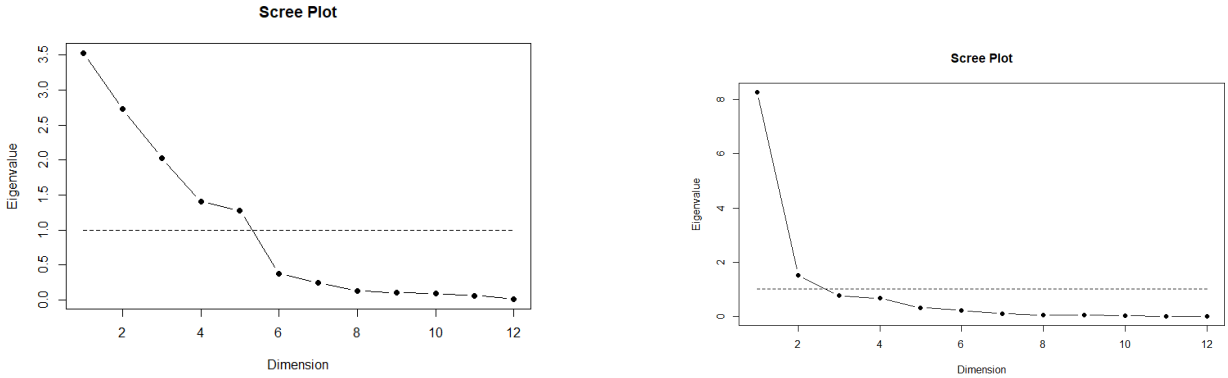
We conduct attribute-based analysis in two steps. We begin our attribute-based study by visually inspecting the main features that define brands' positions on the car perceptual map using PCA. Then, we run our triplet method to compute the raw and adjusted agreement rates between these

describe the term "car brand". More specifically, we use the prompt "The car brand..." on the GPTNeo 2.7B, and manually counted 12 high frequency words that represent attributes of car brands. These 12 attributes are popularity, power, sportiness, style, "technological advanced-ness", eco-friendliness, fuel efficiency, reliability, safety, comfort, durability, and spaciousness.

[3]GPT-4 is asked to rate each brand on each attribute 5 times, and the average is taken as the LLM pairwise rating. We collected 1050 human direct rating responses where each participant is asked to rating 10 brand-attribute pairs

data sets.

As shown by the scree plots on Figure S1, the LLM data set contains 5 significant features and the human data set contains 2; therefore, both data sets contain at least 2 significant principal components. We compare the major features captured by LLM and human data by comparing their first two principal components.

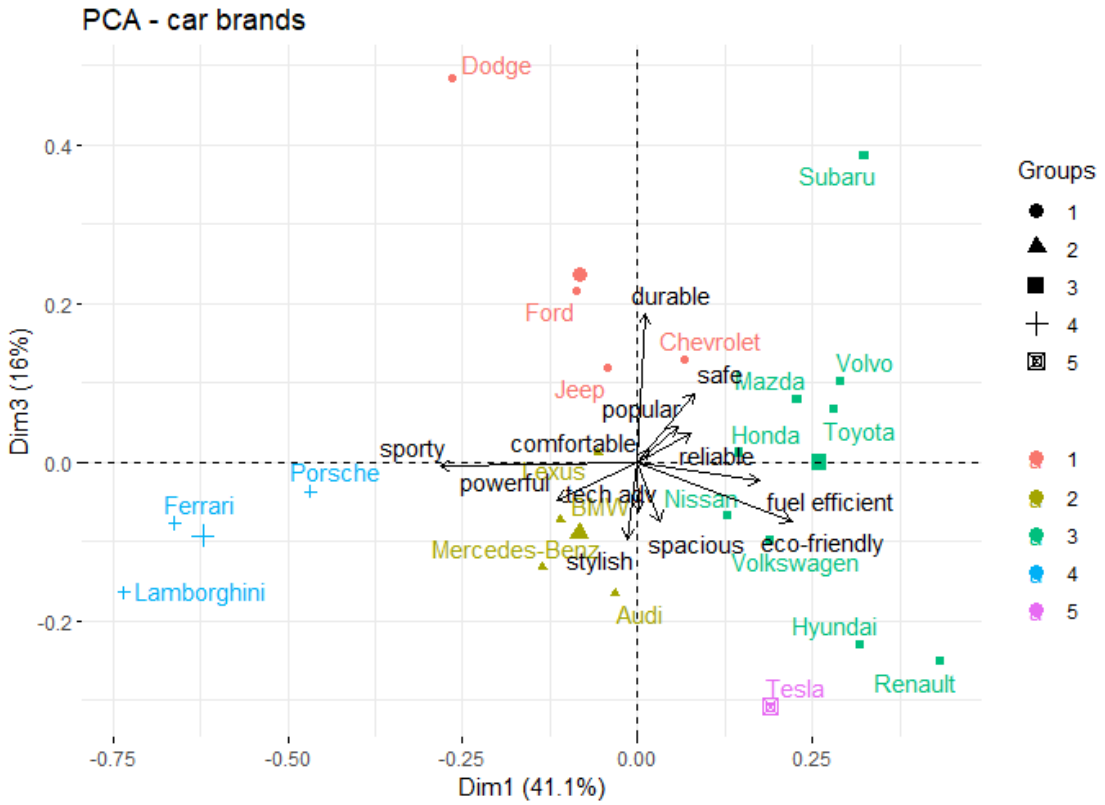


(a) LLM Open-ended (b) Human

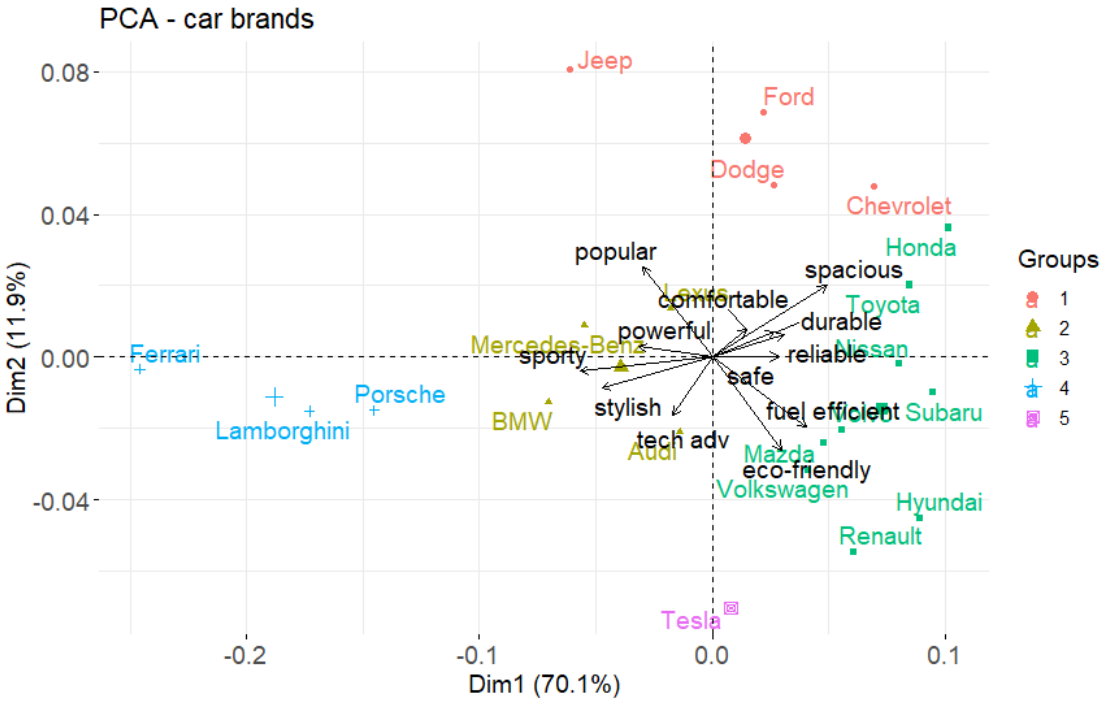
Figure S1: Car study scree plots

Figure S2a and Figure S2b show the first and third component of LLM data and the first two principal components of human data, respectively [4]. Both figures organize 12 attributes into three clusters. Sportiness aligns negatively with the first principal component, relating to power, style, and technology. Safety and comfort attributes align with the positive direction of both principal components, encompassing durability, reliability, etc. The fuel-efficiency cluster aligns positively with the first component and negatively with the second, indicating eco-friendliness and fuel-efficiency. Both maps clusters the 21 car brands into 4 clusters. Sports car brands are on the left side of the map, signifying sportiness and power. Non-sports luxury cars are near the center, signifying well-roundedness. American car brands top both figures, associated with safety and reliability. Mid-range brands locate to the right, linking safety, comfort, and fuel-efficiency. Tesla appears at the bottom, symbolizing a high-tech image. Notably, the key contrast lies in the popularity attribute's positioning. Human ratings tie popularity to sportiness, while LLM responses link it to safety and comfort. This divergence likely emerges from the term "popular" being ambiguous.

[4] In terms of the percent of variation explained by each component, the 2nd and 3rd components are close at 18.6% and 16%. We choose to show the 3rd component because this PCA exercise is mainly for visual inspection and these components are the most similar to the human data



(a) LLM



(b) Human

Figure S2: Car study PCA plots

For an objective comparison, we compute triplet agreement rates between the data sets. Each triplet comprises a brand and two attributes. For example, if both LLM and human data rate Lamborghini's sportiness higher than its safety, it's an instance of dataset-level agreement. Overall agreement between datasets is the proportion of such agreements across all triplets.

Preparing for calculations, we standardize each human data column to account for differing means and variances. For instance, a 7/10 for comfort might differ from a 7/10 for sportiness. To remove such disparities, we standardize attribute ratings across all brands.

Similar to the overall similarity study, we compute a theoretical maximum triplet agreement rate via bootstrapped samples from the same dataset. In table S9, the adjusted agreement rate between LLM open-ended data and human data is 72.1%, while the LLM pairwise numerical rating data's adjusted agreement rate with human data is higher at 74.5%. Notably, even though LLM open-ended data is generated using a less advanced and much smaller model, it performs remarkably well compared to LLM pairwise numerical rating data[5]. This highlights the potential of open-ended prompts for large language models in automating perceptual analysis.

Overall, this study shows that both LLM and human participants can capture the same most important features which are represented by the first two principal components of PCA. Moreover, our results show a high degree of similarity between the two data sets on a triplet level, indicating that we can extract attribute-based insights from open-ended LLM responses.

[5] As of now, it is prohibitively expensive to generate open-ended responses at scale with GPT4 as OpenAI does not provide free access to this model, but we expect similarly capable open source models to be available within a year or two.

Table S9: Triplet agreement rates of car brand attribute-based analysis data sets

	LLM open	LLM pairwise (GPT4)	Human pairwise
LLM open	1 (1,1)	.671 .671	.721 .694
LLM pairwise (GPT4)	–	1 (1,1)	.745 .717
Human pairwise			.925 (.848,.985)

The numbers in black are the point estimates of the self-consistency rates and agreement rates between data sets. The numbers in red are agreement rates between data sets after adjusting for imperfect self-consistency. The numbers in the parentheses are bootstrapped 95% confidence intervals of self-consistency rates. In addition, note that because of the cost constraint we only sample GPT4 5 times, so the bootstrapped self-consistency rates are not meaningful. Here we present conservative estimates by setting them to 1, but realistically, they will not be deterministic and the adjusted agreement rates for GPT4 will be higher.

D Apparel study

In this section, we apply our methodology to a set of 17 major apparel brands (excluding watch and jewelry brands) as shown in Table S10. This list contains a large variety of apparel brands: fast fashion brands, sports and utility brands, and luxury non-sports brands. The list is based on the largest brands by brand value (excluding watch and jewelry brands). Prada, for example, were left off as it is valued at an order of magnitude lower than the leading brand Nike. We use these brands to LLM- and human-generated perceptions of major apparel brands .

Table S10: Table of apparel brands

Brand	Region	Brand value (in billions ($))
Nike	United States	30.4
Gucci	Europe	15.6
Louis Vuitton	Europe	14.9
Adidas	Europe	14.3
Chanel	Europe	13.2
Zara	Europe	13.2
UNIQLO	Asia	13.1
H&M	Europe	12.4
Hermes	Europe	11.7
Dior	Europe	Not disclosed
Coach	United States	Not disclosed
The North Face	United States	Not disclosed
Puma	Europe	Not disclosed
Burberry	Europe	Not disclosed
Ralph Lauren	United States	Not disclosed
Levi's	United States	Not disclosed
Lululemon	Europe	Not disclosed

Note: This list is based on the apparel brand valuation ranking on branddirectory.com. The exact valuation of brands after Hermes is not disclosed. This list contains all non-jewelry and watch brands on the first page of the ranking.

First, we conduct the top-down overall similarity analysis. We collect 10,000 data points for each listed apparel brand by queuing the LLM with prompts "The apparel brand X is similar to the apparel brand...". In addition, we collect open-ended human responses from 501 participants asking them to complete the sentence following the the same prompts, with each participant asked to complete 10 prompts. Similar to the car study, over 95% of the human responses are less than 3

words and the most common type of response has a single brand name. We apply the bag-of-words ordinal embedding method to compute brand pair-level similarity scores. In addition, we collect human numerical similarity ratings from 501 participants, and each participant is asked to rate 10 pairs of brands. We get over 30 ratings for each brand pair. There is a major difference between the car and apparel data sets. In the car direct rating data, the order of the pair is always the same. In other words, all of the participants that are asked to rate the similarity between Mercedes and BMW are given the prompt 'How similar are these two car brands to each other? BMW/Mercedes'. However, in the apparel survey, some participants are asked to rate the similarity between A and B, and the other participants are asked to rate B and A. For example, some participants are given the prompt 'How similar are these two apparel brands to each other? Nike/Adidas', while the other participants are given 'How similar are these two apparel brands to each other? Adidas/Nike'. For each directed rating, we have over 15 ratings. This additional variation in our data provides another opportunity to evaluate the self-consistency of human results. If human perception of brands is self-consistent, simply reversing the order of the two brands should not affect the results. In terms of our data matrix, reversing the order of the brands is equivalent to taking the transpose of the data matrix. Therefore, we have two different versions of human direct rating data which are the transpose of one another. We denote them as human direct rating 1 and human direct rating 2.

As shown in Tables S11 and S12, in both our open-ended and direct rating apparel brand study, the mean and median of the education level of the participants are both within 4 to 5, which corresponds to associated and bachelor degrees. In addition, we observe the middle 80% of participants ranges from education level 2 to level 6, which corresponds to high school graduates and master degree holders. The mean and median of income are both between level 4 and level 5, which corresponds to an annual income of 40 to 50 thousand which is 20% lower than the participants in our studies of car brands. Note that on Prolific.com, participants voluntarily choose which surveys they want to take. Therefore, it is reasonable that participants who are more wealthy self-select to taking the survey about cars, since they may be more familiar with a product like a car that demand more financial investment. Moreover, the distribution of household income is skewed to the right with the first decile at 10 to 20 thousand dollars per year, and the ninth decile is at above 150 thousand dollars per year. This means our data is able to capture some information about high income individuals too. In addition, we observe that the mean and median of response time are both 1 to

2 minutes in the direct rating study, and 2 to 3 minutes in the open-ended study. This agrees with our intuition that it takes time for people to formulate responses when not given a specific set of options. Lastly, we observe that all of participants chose English as their preferred language, so there is no significant language barrier in our study.

Table S11: Human open-ended data descriptive statistics (apparel)

	Education	Income	Gender	Response duration (sec)	Language
Mean	4.14	4.87	0.52	181.97	1
Standard Deviation	1.47	3.04	–	140.98	–
Median	5	4	0	149	1
1st decile	2	2	0	77	1
9th decile	6	10	1	339	1
Min	1	1	0	20	1
Max	8	12	3	2197	1

The language indicator is 1 if the language is English, and it is 0 for all other languages. In the gender column, 0 stands for male, 1 stands for female, 2 stands for non-binary, and 3 stands for rather not say.

Table S12: Human pairwise data descriptive statistics (apparel)

	Education	Income	Gender	Age	Response duration (sec)	Language
Mean	4.02	4.91	0.52	36.26	102.43	1
Standard Deviation	1.46	3.00	–	13.15	70.74	–
Median	5	4	0	33	86	1
1st decile	2	2	0	22	55	1
9th decile	6	10	1	56	154	1
Min	1	1	0	18	29	1
Max	8	12	3	89	959	1

The language indicator is 1 if the language is English, and it is 0 for all other languages. In the gender column, 0 stands for male, 1 stands for female, 2 stands for non-binary, and 3 stands for rather not say.

Table S13: Triplet agreement rates of apparel brand data sets

	LLM open	LLM pairwise (GPT4)	Human open	Human pairwise 1	Human pairwise 2 (transpose)
LLM open	.964 (.954,.976)	.662 .650	.782 .745	.730 .658	.734 .665
LLM pairwise (GPT4)	–	1 [1,1]	.781 .758	.872 .803	.839 .775
Human open	–	–	.941 (.917,.964)	.923 .829	.904 .808
Human pairwise 1	–	–	–	.842 (.823,.861)	.981 .829
Human pairwise 2	–	–	–	–	0.847 (.822,.873)

Black numbers show the point estimates of the self-consistency rates and agreement rates between data sets. Red numbers are agreement rates between data sets after adjusting for imperfect self-consistency. The numbers in the parentheses are bootstrapped 95% confidence intervals of self-consistency rates. In addition, note that because of the cost constraint we only sample GPT4 five times, so the bootstrapped self-consistency rates are not meaningful. Here we present conservative estimates by setting them to 1, but realistically, they will not be deterministic and the adjusted agreement rates for GPT4 will be higher.

These agreement rates highlighted in red in Table S13 are agreement rates between different data sets. More specifically, when comparing the LLM open-ended data set with the human pairwise numerical rating data set 1, we observe a 65.8% agreement rate which is 73.0% of the empirical maximum of 90.3%, computed using property 2. Similarly, the agreement rate between the LLM open-ended data set and human direct rating 2 is 73.4% of the empirical maximum of 90.6%. The agreement rate of the LLM open-ended data set versus human open-ended responses is 78.2% of the empirical maximum of 95.3%. The adjusted agreement rate between the LLM pairwise numerical rating data set and the human data sets are even higher, ranging from 78.1% to 87.2%. When comparing human open-ended responses to human pairwise numerical ratings, we also should not expect the agreement rate to be higher than the average of the two self-consistency rates. Therefore, the adjusted agreement rates between human open-ended and human direct rating are 92.3% (versus direct rating 1) and 90.4% (versus direct rating 2) of the empirical maxima of 89.1% and 89.4% respectively. Moreover, even though numerical evaluations are not native to large language models,

a much larger model like GPT4 that is asked to directly give numerical ratings can still outperform a smaller GPTNeo which is asked to give open-ended responses. This suggests that larger and more advanced language models are more adept at interpreting brand similarities.

In addition, we closely examine the source of inconsistency within the human pairwise numerical rating data set. At first glance, the agreement rate between the two versions of the direct rating data set is only 82.9%, which is lower than the self-consistency rates of LLM and human open-ended data sets by more than 11%. However, when compared to the self-consistency rate of the two direct rating data sets, the agreement rate between the two is 98.1% of the empirical maximum of 84.5%. Therefore, we infer that most of this inconsistency of results that appears when the order of the two paired brands is reversed is due to the innate self inconsistency of the human direct rating data, not the order reversal.

In addition to the top-down approach, we show that our method also shows a high level of similarity between LLM and human data sets using the bottom-up attribute-based approach. More specifically, we try to find the main features that define apparel brands' positions on the perceptual map using data collected from an LLM model. To do this, we collect 6000 LLM responses and 1001 human responses. The LLM responses are sentence completions of open-ended prompts such as "The most A apparel brand is...", where A is an attribute of apparel brands.[6] In addition, to replicate the LLM pairwise numerical rating data in the overall similarity study, we ask GPT-4 to directly rate each brand on this set of attributes. We repeat each question 5 times and take the average as the LLM rating. For human data, we ask each participant to give a numerical rating from 0 to 10 on answering the prompt "How A is the apparel brand X...". Attribute-based analysis is done in two steps. First, we use Principal Component Analysis to group attributes into features and map brands according to the first two principal components. Then, we run our triplet method to compute the agreement rates between data sets.

Table S14 shows the summary statistics of the human data for attribute-based analysis of apparel brands. Both the mean and median household income are between 60,000 to 70,000 which is representative of the average household income of the United States. We have roughly equal

[6] In this study, we include 14 attributes that appeared often when we ran the prompt "The apparel brand..." on the LLM. These 14 attributes are popularity, how well-known the brand is, recognizability, comfort, sportiness, how athletically focused the brand is, sportiness, trendiness, technology, creativity, how high-end the brand is, luxury, exclusivity, and how fashionable the brand is.

numbers of male and female participants in our study. The middle 80% of participants are in between 21 and 56 years old. In addition, the average response time of each participant is just over a minute. Lastly, all participants chose English as the user's preferred language, which implies there is no significant language barrier in our study.

Table S14: Human attributed-based analysis data descriptive statistics (apparel)

	Income	Gender	Age	Response duration (sec)	Language
Mean	6.51	0.56	35.59	65.39	1
Standard Deviation	3.50	–	13.40	70.99	–
Median	6	1	32	49	1
1st decile	2	0	21	28	1
9th decile	11	1	56	103	1
Min	1	0	18	11	1
Max	12	3	80	1208	1

The language indicator is 1 if the language is English, and it is 0 for all other languages. In the gender column, 0 stands for male, 1 stands for female, 2 stands for non-binary, and 3 stands for rather not say.

First, we observe that some responses in the open-ended data are about brands outside of our focus set. Therefore, similar to the car study, we scale each column (all responses under one prompt), so we only account for responses that are about brands in our focus set. In addition, we standardize each column of the pairwise numerical rating data sets to avoid the problem described in the car attribute-based study. Next, we quantify the similarity between these data sets using the triplet method. As shown in Table S15, both the LLM open-ended and human data sets have high self-consistency rates at 91.8% and 96.4% respectively. Moreover, the agreement rate between the two data sets is 67.8% compared to the empirical maximum of 94.2%. More impressively, the adjusted agreement rate between human direct rating and GPT-4 based LLM pairwise numerical rating results is 73.3%.

Table S15: Triplet agreement rates of apparel brand attribute-based analysis data sets

	LLM Open-ended	LLM Direct Rating (ChatGPT-4)	Direct rating
LLM Open-ended	.918 (.846,.978)	.601 .576	.678 .638
LLM Direct Rating (ChatGPT-4)	–	1 [1,1]	.733 .720
Direct rating	–	–	.965 (.912,1)

The numbers in black are the point estimates of the self-consistency rates and agreement rates between data sets. The numbers in red are agreement rates between data sets after adjusting for imperfect self-consistency. The numbers in the parentheses are bootstrapped 95% confidence intervals of self-consistency rates. In addition, note that because of the cost constraint we only sample GPT4 five times, so the bootstrapped self-consistency rates are not meaningful. Here we present conservative estimates by setting them to 1, but realistically, they will not be deterministic and the adjusted agreement rates for GPT4 will be higher.

E Heterogeneity

In this section, we present evidence that there is no difference in different demographics' perception of car brands. In addition, we define the regressions that show the variations in different demographics' preferences of car brands.

Focusing on the data collected for cars, we start by showing that consumer heterogeneity does not play a significant role when we study brand perceptions. In our case studies, we distinguish consumers by their age, income, and gender. Each consumer is defined by an indicator triple – an old, poor, male is defined as $(0, 0, 0)$, and a young, rich, female is defined as $(1, 1, 1)$. Survey respondents are given the opportunity to choose "other" as their gender, but because the portion of respondents who chose "other" is infinitesimal, we drop these consumers instead of making a separate category. For income and age, we drop all consumers within a middle sub-range between the maximum and minimum. This creates a discontinuity that separates rich from poor and old from young. More specifically, we define young as at or below 40 years old and old as at or above 50; we define poor as annual income below $40,000 and rich as annual income above $100,000.[7]

We first examine consumer heterogeneity's effect on consumers' perception of a brands using the human data. We reuse the brand similarity data we collected for our car brand case study, and divide this data set into 8 subsets each defined by a different indicator triple. Using the triplet method, we can measure the similarity between pairs of data matrices where two out of three covariates are controlled. We get at least 64.7% and up to 95.4% agreement rate between different data sets. This implies that in the human data, consumer heterogeneity has limited impact on consumers' perception of car brands (see Table S17).

To check whether GPTNeo results can lead to the same conclusion, we collect new data using prompts like "A young and poor male thinks the car brand X is similar to the car brand..." versus "A young and poor female thinks the car brand X is similar to the car brand...", etc. We find that the brand similarity matrices are highly similar when the LLM is asked to complete sentences from different perspectives – male versus female, young versus old, and rich versus poor (see Table S16). As expected, we do not find meaningful differences between different groups when examining consumer perceptions.

[7] We take a conservative estimate of the top and bottom 40% of U.S. household income.

	old and rich	old and poor	young and rich	young and poor
male vs female	.906	.899	.887	.887
	old and male	old and female	young and male	young and female
rich vs poor	.900	.890	.869	.860
	poor male	poor female	rich male	rich female
young vs old	.881	.891	.882	.862

Table S16: This table presents the similarities between pairs of car data matrices using the LLM data. Two out of three of age, gender, and income are controlled in each pair.

	old and rich	old and poor	young and rich	young and poor
male vs female	.720	.647	.767	.720
	old and male	old and female	young and male	young and female
rich vs poor	.954	.953	.857	.867
	poor male	poor female	rich male	rich female
young vs old	.678	.682	.712	.663

Table S17: This table presents the similarities between pairs of car data matrices using the human direct rating data. Two out of three of age, gender, and income are controlled in each pair.

However, we use linear regressions to find that there are variations in brand preferences of different demographics. More specifically, we run the following linear regressions to test whether age, income, or gender have significant correlations with consumers' preference for expensive cars, sports cars, and family cars.

$$P(\text{expensive}|(\mathbb{1}_{age}, \mathbb{1}_{income}, \mathbb{1}_{gender})) = c_{\text{expns}} + c_{\text{age}}^{\text{expns car}}\mathbb{1}_{age} + c_{\text{income}}^{\text{expns car}}\mathbb{1}_{income} + c_{\text{gender}}^{\text{expensive car}}\mathbb{1}_{gender}$$

$$P(\text{sports}|(\mathbb{1}_{age}, \mathbb{1}_{income}, \mathbb{1}_{gender})) = c_{\text{sports}} + c_{\text{age}}^{\text{sports car}}\mathbb{1}_{age} + c_{\text{income}}^{\text{sports car}}\mathbb{1}_{income} + c_{\text{gender}}^{\text{sports car}}\mathbb{1}_{gender}$$

$$P(\text{family}|(\mathbb{1}_{age}, \mathbb{1}_{income}, \mathbb{1}_{gender})) = c_{\text{family}} + c_{\text{age}}^{\text{family car}}\mathbb{1}_{age} + c_{\text{income}}^{\text{family car}}\mathbb{1}_{income} + c_{\text{gender}}^{\text{family car}}\mathbb{1}_{gender}$$[8].

Using these regressions, we find that in both LLM and human data sets consumers' income

[8]Readers may suspect that there is a multicollinearity problem between age and income. We eliminate this concern in the data collection process by collecting balanced samples. We collect roughly equal amounts of old rich and young poor samples and vice versa

level significantly affects their preference of car brands. Wealthy consumers prefer sports cars and expensive cars while poor consumers prefer family cars. In addition, we find that males like expensive cars more than females. The exact coefficients are shown in table S18.

Table S18: Consumer heterogeneity on brand preferences

	Expensive	Sporty	Family
Age (LLM)	0.026***(0.0031)	0.021***(0.0019)	-0.12***(0.0043)
Age (Human)	0.047 (0.032)	-0.0068 (0.0095)	-0.0043 (0.035)
Income (LLM)	0.11***(0.0031)	0.059***(0.0019)	-0.18***(0.0043)
Income (Human)	0.16***(0.032)	0.034***(0.0095)	-0.15***(0.036)
Gender (LLM)	-0.0023*** (0.0031)	-0.019*** (0.0019)	-0.0041(0.0043)
Gender (Human)	-0.082**(0.032)	-0.012 (0.0095)	0.075**(0.035)
Const. (LLM)	0.083***(0.0031)	0.017*** (0.0019)	0.51***(0.0043)
Const. (Human)	0.17***(0.032)	0.0095 (0.0093)	0.73***(0.035)

F Prompt selection

Multiple prompts are tested in the prompt selection section. We provide the exact prompts test in this part of the supplemental materials:

Simple prompts:

On a scale of 0 to 10, how similar are the car brands A and B?

Few-shot prompts:

Question: on a scale of 0-10, how similar are the car brands a_1 and a_2 on a scale of 0 to 10 where 10 means very similar?

Answer: 5

Question: on a scale of 0-10, how similar are the car brands a_3 and a_4 on a scale of 0 to 10 where 10 means very similar?

Answer: 1

Question: on a scale of 0-10, how similar are the car brands a_5 and a_6 on a scale of 0 to 10 where 10 means very similar?

Answer: 3

Question: on a scale of 0-10, how similar are the car brands a and b on a scale of 0 to 10 where 10 means very similar?

Answer:

RTF (role, task, format) prompts:

I want you to act as a person filling out a survey. 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?

Combined prompts:

I want you to act as a person filling out a survey. 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.

Question: on a scale of 0-10, how similar are the car brands a_1 and a_2 on a scale of 0 to 10 where 10 means very similar?

Answer: 5

Question: on a scale of 0-10, how similar are the car brands a_3 and a_4 on a scale of 0 to 10 where 10 means very similar?

Answer: 1

Question: on a scale of 0-10, how similar are the car brands a_5 and a_6 on a scale of 0 to 10 where 10 means very similar?

Answer: 3

Question: on a scale of 0-10, how similar are the car brands a and b on a scale of 0 to 10 where 10 means very similar?

Answer:

G Applicability to different product categories

In this section we investigate if the LLM-powered perceptual analysis method performs differently in product categories with different levels of involvement. More specifically, we compare these agreement rate (overall similarity) across three categories: car, apparel, and hotel. These three categories are ranked in decreasing order of involvement. Most people drive the same car for at least a few years after a purchase, most people frequently wear a piece of apparel for at least a few weeks after a purchase, and most people only live in the same hotel for a few days.

Data collection details for car and apparel brands are specified in the main paper and supplemental materials section C and D. Here we describe the data collection process for hotel brands in more detail. First, similar to the car and apparel studies, we sampled GPTNeo-2.7B with the prompt "The hotel brand" and choose the brands with a high frequency when intersected the list of hotel brands by U.S. News. We included 15 hotels in total. Five star hotels including Four Seasons, Grand Hyatt, Ritz-Carlton, Fairmont, W Hotel, and JW Marriot. Three or four star hotels like Sheraton, Courtyard, Hyatt Regency, Hilton, Marriot, and Westin. Business hotels like Best

Western, Holiday Inn, and Residence Inn. For LLM data collection, we collect 15000 open-ended responses using GPTNeo following the prompt: "The hotel brand X is similar to the hotel brand". For human responses, we collect 499 open-ended human survey responses on *Prolific.com* using the same prompt, and we also collect 500 direct rating responses where we direct ask each participant to rate the level of similarity of 10 pairs of hotel brands on a scale from 0 to 10. Data set level raw agreement rates are calculated using the triplet method described in section B of supplemental materials.

Table S19: Raw agreement rates in different product categories

	Car	Apparel	Hotel
LLM vs Human Open-ended	.747	.745	.626
LLM vs Human direct rating	.720	.658	.573

As shown in table S19, the raw agreement rates for car brands are higher than apparel brands which are higher than hotel brands. This suggests that the applicability of LLM-powered perceptual analysis might be correlated with the level of involvement consumer has with the product after a purchase. Intuitively, if a consumer has a higher degree of involvement with a product post purchasing, he or she may take more careful considerations when purchasing a product, and therefore, the perception of these brands may be more consistently represented in the training corpus of LLMs ("the pile" in the case of GPTNeo). However, the careful examination of this observation and intuition is an important direction for future studies.

H LLM responses with multiple brand mentions

One extension to our study is to consider open-ended LLM responses where more than one brands are mentioned, like "The car brand A is similar to the car brand B and also C". Here, we test whether the order of brands mentioned in this case carries any information: brand A is more similar to B than C. This is a natural application for our triplet method. It is worth noting that in our entire car open-ended data set generated by GPTNeo, there are only 1384 out of 357000 responses (roughly 0.4%) that contain more than one brands, and only 13 with more than 2. This is partly because we limit the output token length to at most 20; as a result, most responses only contain one complete sentence. Nonetheless, there is at least one such response to each of the 21 prompts (one for each brand). Since the number of responses containing more than 2 brands is negligible, we only focus on the first two brands mentioned after the prompt excluding the brand in the prompt.

We compare these 1384 triplets with human open-ended survey data and human direct rating survey data using the triplet method. We find that they agree with the open-ended data set 47.3% of the time and agree with the direct rating data set 55.9% of the time. This seems to imply that in the case where the LLM's response contains more than one brands, the order of brands does not carry much implication for similarity ranking. Note that we only consider a case where there are limited occurrences of LLM responses containing multiple brands. We leave the more complete validation of the general case with longer responses to future research.

I Comments on LLM generated data

In this section, we present some comments regarding the high self-consistency rate of LLM generated open-ended data and our thoughts on the importance of including a year specification in the prompt.

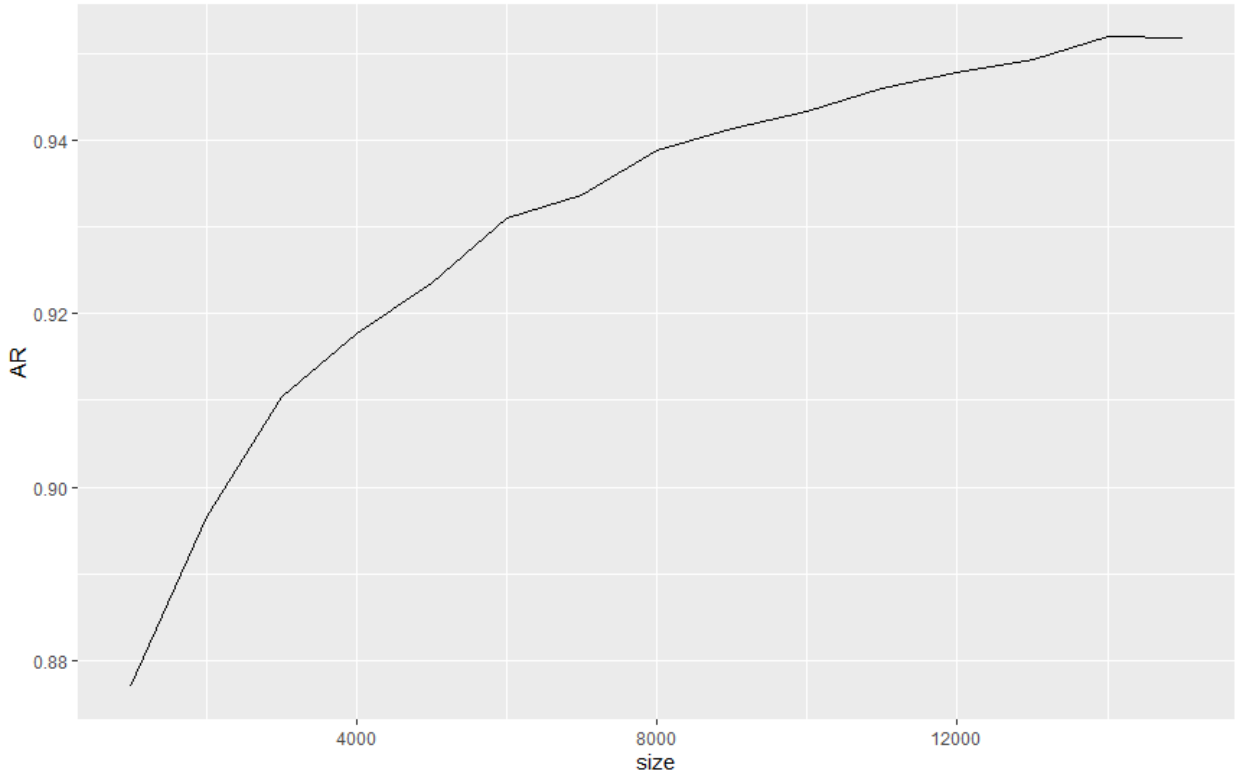


Figure S3: self-consistency rate of LLM open-ended data as the number of samples increases

The self-consistency rate of LLM open-ended data is dependent on the number of samples we ask the LLM to generate. In figure S3, we show that when we only take a small sample of 1000 open-ended responses, the self-consistency rate is below 90%, but as the sample sizes increases, the LLM data set becomes more self-consistent. This is an intuitive result, and this is one of the advantages of using an LLM model: it is feasible in many cases to sample enough data points such that the self-consistency rate is near 100%.

In regard to adding a time (year) specification in the prompt, we speculate that this likely causes LLMs to put more weight on its training data close to this time period. One suggestive evidence is that when the financial market experienced a drastic shock around 2008, the prompt with a year specification is much more similar to the car switching behavior of human compared to the prompt without any year specification. We cannot draw a conclusion purely based on this observation, and we leave the more formal validation of this to future research.

J Perceptual maps over time

In this section, we plot the perceptual maps over time (1999,2002,2005,2008) using the car trade-in data and the LLM pairwise numerical rating data with year specification. As shown in figures S4 and S5, we observe that the maps do not change very drastically over the period of 1999 to 2005, and between 2005 and 2008, there is a more noticeable change especially more higher-end car brands. It is worth noting that we should use the maps as qualitative and suggestive evidence. For a more observative analysis, please refer to section 3.2 in the main paper.

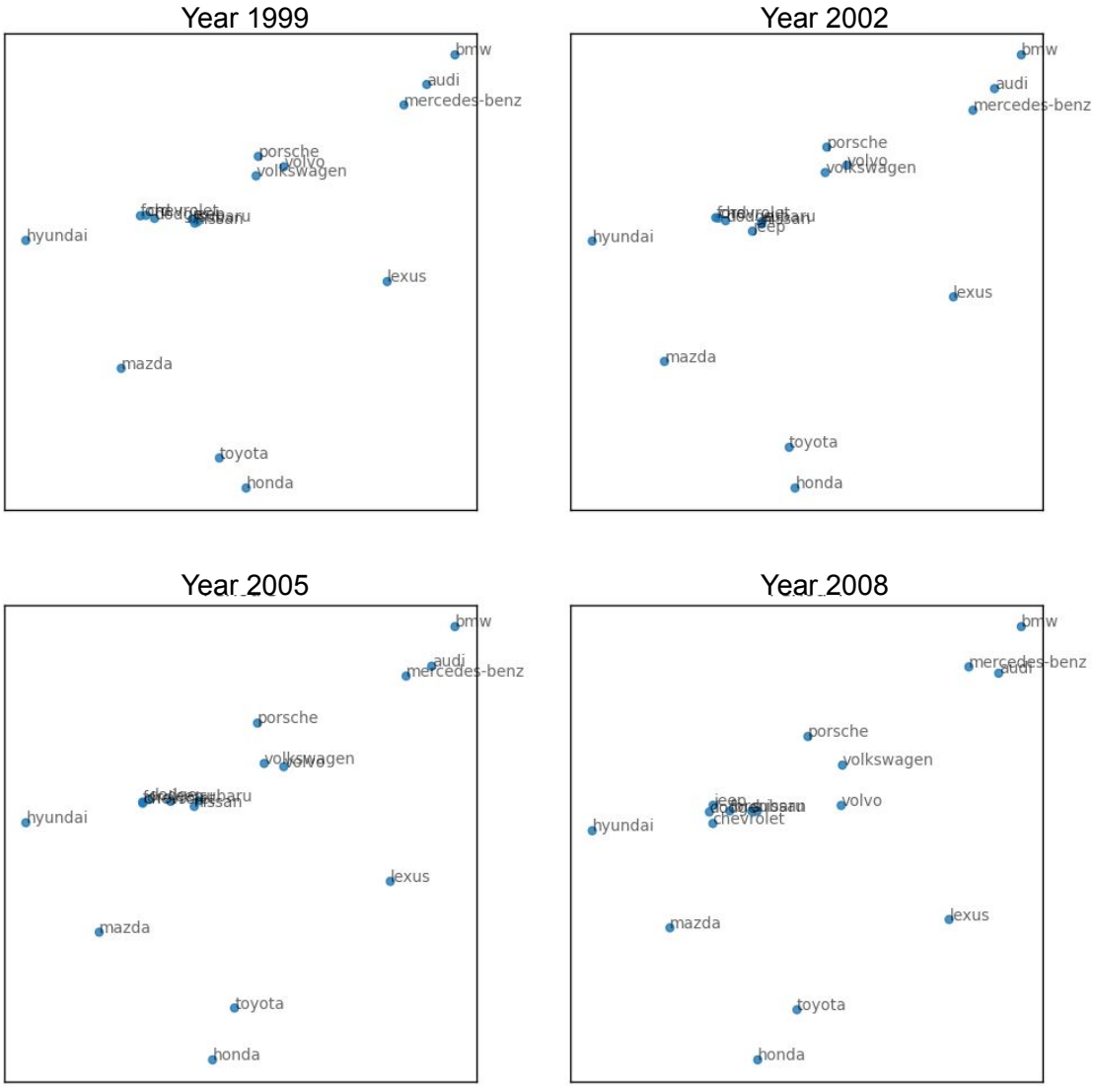


Figure S4: Perceptual maps using LLM pairwise data with year specification

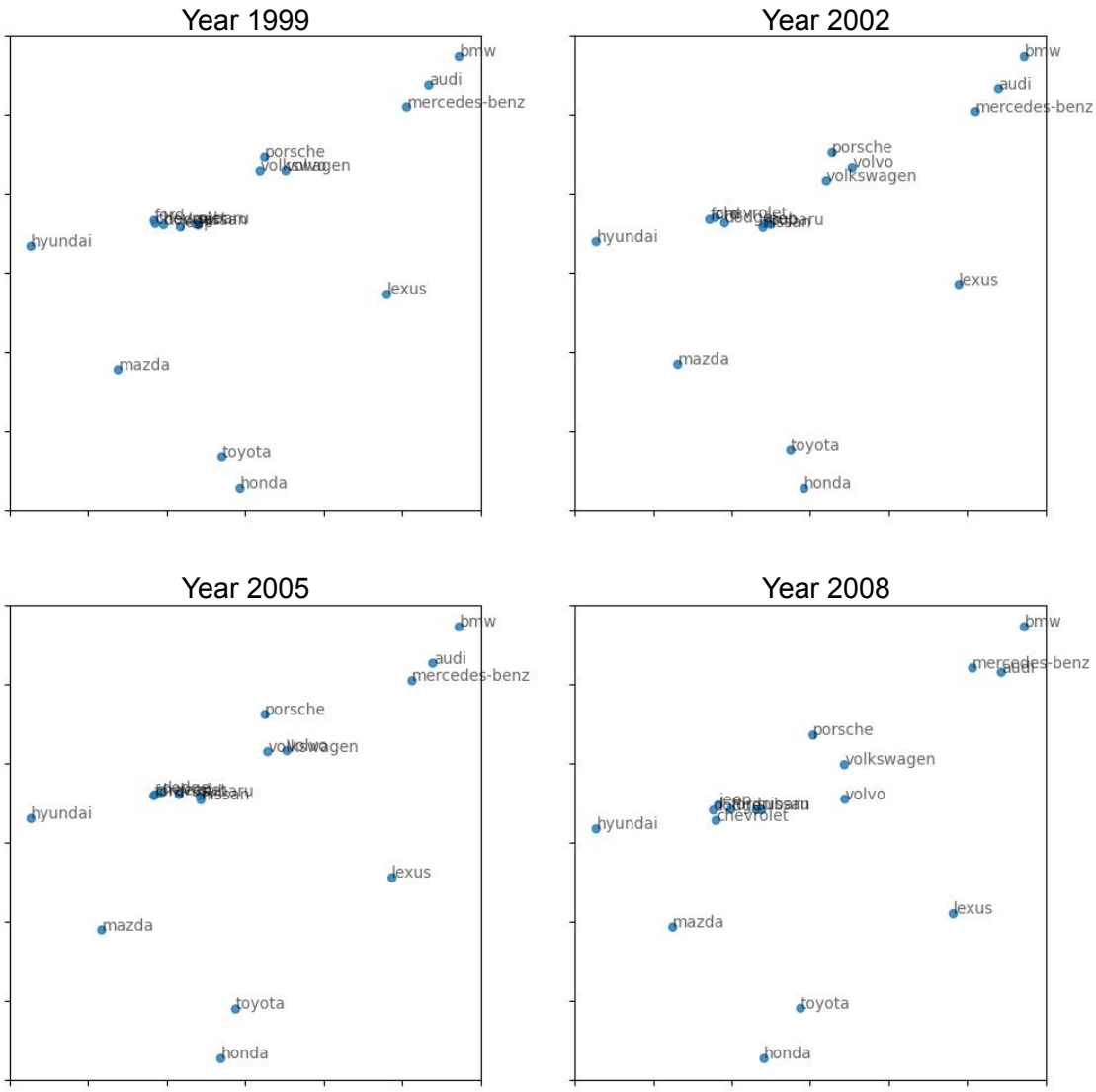


Figure S5: Perceptual maps using car trade-in data by year