UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Analyzing Brand Perception In LLMs

Author:
Jaime Leonardo SÁNCHEZ SALAZAR

Supervisor:
Dr. Oriol PUJOL VILA
Dr. Santiago SEGUÍ MESQUIADA

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

June 30, 2024

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Analyzing Brand Perception In LLMs

by Jaime Leonardo SÁNCHEZ SALAZAR

This thesis investigates brand perception in different Large Language Models (LLMs), focusing on three brands: Apple, Samsung, and Huawei. We first established an understanding of brand perception and the construction of psychometrically sound tests. Leveraging this foundation, we defined four metrics across two dimensions, sentiment and preference, to facilitate a comprehensive analysis. In the sentiment dimension, we observed that the Gemma LLM exhibited consistent bias across all brands, whereas ChatGPT3.5 and ChatGPT4 displayed similar behavior for Apple and Samsung, with notable differences for Huawei. In the preference dimension, all studied LLMs demonstrated transitivity consistency, consistently preferring Apple over Samsung and Samsung over Huawei. Our findings highlight the potential for extensive analysis using the defined metrics, limited here by time constraints. We suggest several avenues for future research, including expanding the range of brands and LLMs analyzed, improving the question bank through collaboration with psychologists, and incorporating varied question connotations and mask questions to enrich the study's depth. This study provides a methodological framework for assessing brand perception in LLMs, with implications for broader applications beyond the specific brands and models examined.

Acknowledgements

Firstly, I would like to thank the supervisors of this project, Dr. Oriol Pujol Vila and Dr. Santiago Seguí Mesquiada, for their guidance whenever needed in shaping this work. Your time, effort, and advice have been crucial in successfully completing this project. I would also like to extend special thanks to Dr. Antonio Solanas Pérez, Dean of the Faculty of Psychology, for assisting us in structuring parts of this work.

To my family, thank you for your constant support and understanding throughout this journey. Your patience and encouragement have been incredibly important to me, and I would not be who I am without it. From the bottom of my heart, thank you.

To my classmates, your collaboration and friendship have been fundamental in making this year much more enjoyable and engaging. Special thanks to David and Madison, very good friends of mine.

Contents

Chapter 1

Introduction

1.1 Motivation

The use of Large Language Models (LLMs) have surged with technology developments and adoption by industry in recent years. Their ability to generate coherent and contextually relevant text has led to their use in a wide range of applications, from virtual assistants to automated content generation.

As the practice, trust, and dependence on such models progresses, inherent biases in language models arise. LLMs can perpetuate and amplify existing biases, raising significant ethical concerns. This can lead to unfair representations of individuals or groups, negatively impacting decisions and perceptions. It is crucial to address these biases to ensure that artificial intelligence technologies are fair and equitable.

But what relationships exist between LLMs and the perception of a brand? And why is this important to study? LLMs are increasingly integrated into platforms that interact directly with consumers, such as chatbots, virtual assistants or recommender systems. These models interpret and generate text-based data that reflect public perceptions and sentiments towards a brand. By processing vast amounts of text data from various sources, such as social media, reviews, and surveys, to identify trends, opinions, and attitudes they can influence public opinion and perception of those brands.

Additionaly, ensuring these models are transparent and trustworthy is essential as consumers and businesses begin to rely more on LLMs for information and decision making. Not only that, but also understanding and measuring how LLMs perceive and represent different brands is crucial to maintaining the integrity and reliability of these technologies.

Bias in LLMs also has commercial implications: companies invest significantly in building and maintaining their brand image. Negative bias in an LLM could undermine these efforts, while positive bias could create an unfair advantage. Bearing this in mind, companies must understand how their brands are perceived by such technologies.

Identifying biases and creating metrics to measure brand perception in LLMs can drive continuous improvements in the development of these models. This will not only benefit brands, but also contribute to the overall advancement of artificial intelligence technology, promoting fairer models with less bias and more equity.

1.2 Objectives

For these reasons, our primary objective in this thesis is as follows: *to develop a set of metrics for evaluating brand perception in LLMs*. This leads us to define more specific objectives:

1. Study the bias that an LLM exhibits towards a brand.
2. Propose a set of metrics that can be used to quantify this bias.
3. Propose a set of questions, forming a suitable test, to query the LLM.
4. Measure the aforementioned metrics and apply them to different LLMs and various brands.

1.3 Results

The results of this thesis are:

1. We have defined four metrics, distinguishing between two dimensions, sentiment and preference. From this, the bias of an LLM towards a brand can be studied.
2. A question set is defined in order to test LLMS. Moreover, different formulations of these questions have been developed to adapt them to various question formats.
3. Three different LLMs Gemma, ChatGPT-3.5, and ChatGPT-4 were tested using the aforementioned metrics. The brand perception of Apple, Samsung, and Huawei was evaluated.
4. As a result of the manner in which the study was conducted, the methodology explained here is easily applicable to the study of LLMs not only in relation to brands, but also to other concepts.

1.4 Report Layout

Regarding the structure of this work, it consists of three main parts. The first chapter serves as an introduction in which we explain the main components of this thesis: LLMs and brand perception. The final part of the first chapter is dedicated to the creation of a test from a psychometric point of view and to discussing the existing literature on this topic.

Next, we outline the methodology followed and define a set of metrics used to help us characterize the perception an LLM has of a particular brand. This is followed by the experimental part of the work in which we apply these metrics to specific brands. This will later be used for a comparative analysis. From this analysis, we can draw conclusions about our metrics and, consequently, about the brands and LLMs used.
All the code developed for this research can be found in the following GitHub repository.

https://github.com/jshz12/Analyzing-brands-in-LLMs-Master-Thesis-

Chapter 2

Background and State of the Art

In this section, we aim to familiarize the reader with the concepts addressed throughout this thesis. We briefly explain Large Language Models and introduce the concept of *brand perception*. Following this introduction, we discuss the existing literature on our subject, a crucial foundation for the commencement of our research.

2.1 What are LLMs?

LLMs are a type of Artificial Intelligence (AI) trained to understand, generate, and interact with human language in a way that is both coherent and contextually relevant. These models are 'large' not only in their size, spanning billions of parameters, but also in the vast amount of data they are trained on. This training involves the analysis of a wide array of text sources, from books and articles to websites and social media posts (Czerny, 2024).

The core technology behind LLMs is what is known as a transformer model, first introduced in a 2017 paper titled "Attention Is All You Need". This model uses attention and self-attention to weigh the importance of different words in a sentence. By doing so, it captures the nuances of language, including context, tone, and syntax. The training process involves feeding the model a large corpus of text and using machine learning algorithms to adjust the model's parameters to predict the next word in a sentence.

LLM architecture consists of multiple layers of neural networks. These layers include embedding, recurrent, layers, and attention layers. Each layer helps the model process the input text and generate output predictions (Kesrwani, 2023).

The **embedding layer** converts each word in the input text into a high-dimensional vector representation. This representation captures semantic and syntactic information about the word, which helps the model understand the context.

The **feedforward layers** apply non-linear transformations to the input embeddings. This helps the model learn higher-level abstractions from the input text.

The **recurrent layers** interpret information from the input text in a sequence. They maintain a hidden state that is updated at each time step, allowing the model to capture the dependencies between words in a sentence.

The **attention mechanism** allows the model to focus selectively on different parts of the input text. This helps the model attend to the input text's most relevant parts

and generate more accurate predictions. For instance, when translating a sentence, the model might focus more on the subject of the sentence when translating a verb.

In summary, the architecture of LLMs is designed to process the input text in a way that captures the meaning of the text and the relationships between words. LLMs are known to be useful for a wide range of applications, for instance:

1. **Content Creation:** From writing articles and poems to generating creative fiction, LLMs can produce diverse forms of written content.
2. **Language Translation:** LLMs can translate text between various languages with high accuracy.
3. **Chatbots and Virtual Assistants:** LLMs power sophisticated chatbots that can handle complex customer service inquiries or provide companionship.
4. **Information Extraction and Analysis:** They are used to extract information from large datasets, summarize texts, and even analyze sentiment.
5. **Educational Tools:** In education, LLMs assist in creating personalized learning materials and tutoring systems.

2.2 Brand Perception

Brand perception, often called brand image (Latif et al., 2016) refers to customers' perceptions and associations of a brand stored in their memory. These perceptions shape customers' overall impressions of the brand and influence their emotional responses. A strong brand image helps differentiate a brand from its competitors in the marketplace, conveying a superior message. This, in turn, impacts customers' behavior and purchasing decisions, as they tend to favor brands with a positive image associated with quality and value. Brand image also plays a crucial role in fostering outcomes such as brand familiarity, customer satisfaction, trust, and attitudinal brand loyalty, sustaining its effects over an extended period in customers' minds.

Therefore, brand perception is shaped by various factors that encompass the overall impression a brand leaves on its audience. These factors are influenced by direct and indirect interactions such as:

1. **Customer Experience**: Includes every interaction a consumer has with a brand, from the initial point of contact to post-purchase support.
2. **Trust and Reputation**: Foundational to building long-term relationships with consumers. A brand that is perceived as trustworthy and reputable is more likely to retain current customers and attract new ones through word-of-mouth.
3. **Product Quality**: Directly influences consumer satisfaction and brand loyalty. High-quality products and services reinforce a positive brand image, while poor quality can lead to negative reviews and decreased sales.
4. **Emotional Connection**: Can drive deeper loyalty and advocacy. Brands that resonate emotionally with consumers can create strong, enduring relationships that go beyond transactional interactions.

2.3 Psychometric Analysis

When creating a test, it is essential to develop an appropriate prompt to evaluate an LLM, or any individual. In psychometrics (Renom, 1992), there are three key characteristics that are essential for a test to be effective and from which meaningful conclusions can be drawn. These include:

1. **Unidimensionality**: This characteristic means that the test measures a single dimension or construct. In other words, the items on the test are designed to assess one specific trait, skill, or psychological characteristic. To assess such unidimensionality, different methods are used, such as factorial analysis, which examines the correlation among a set of variables (test questions) to uncover any latent structure.

2. **Reliability**: Reliability refers to the consistency of the scores obtained from the test. A reliable test produces similar results under consistent conditions over time. Two common ways to measure this characteristic are: through the *test-retest method*, where the test is administered to the same group of individuals at two different points in time, and then the correlation between the scores obtained on both occasions is calculated. Or using *Cronbach's alpha coefficient*, which calculates the average correlation among all items of the test and provides a measure of the test's internal consistency.

3. **Validity**: Validity is the extent to which a test measures what it claims to measure. That is, it indicates the precision with which the scores obtained in a test reflect the specific characteristic that is being measured. Here, we can identify two[1] types of validity: *content validity* and *criterion validity*. The first one ensures that the test items adequately represent the entire domain of interest. It is often established through expert judgment and a thorough review of the test items against the content domain.

   On the other hand, criterion validity distinguishes between two types. Concurrent validity, where the test responses are compared with results obtained from another test that is already established and recognized as valid for measuring the same characteristics. While the second one, predictive validity, refers to the test's ability to predict future performance in a specific task or situation.

That being said, the proper creation of a test is more difficult than it seems, as it requires meeting a set of essential characteristics that ensure its quality and usefulness.

2.4 State Of The Art

Do LLMs store implicit associations between brands and brand image attributes? Do the associations embedded in LLMs signify any bias? Do LLMs capture the brand personality intended by a brand? All these questions are useful in order to understand the brand perception that an LLM has towards a brand. All these questions are answered in Srivastava et al., 2021 where they study brand perception through a set of dimensions/adjectives attributed to a brand. When we talk about bias, we

[1] There is, in fact, another last type of validity called construct validity. We do not explain this one here because it is a bit more complex.

quickly encounter a plethora of literature on bias focused on the discrimination of minority groups in LLMs. For our study, this aspect will also be very helpful, as the methodology used is important. Specifically, in Scherrer et al., 2024 they explore the capabilities and implications of LLMs in understanding moral decision-making. Let us briefly introduce each of these papers.

The paper Srivastava et al., 2021 delves into the ability of LLMs to capture brand perception and affect associations. To study this factor, the focus is on how LLMs perceive the relationship between brand names and a set of defining attributes. These attributes, they assert, are a set of adjectives that align more or less with the brand in question. For example, whether the shoe brand Converse is "stylish" or not.

The paper conducts different types of experiments, highlighting the "affect similarity" metric, which captures how similarly a brand and an attribute are represented in the vector space of a language model's layer through cosine similarity. They compute this metric not only between a brand word and different attributes, but also between different brands and different attributes. That is, they compare how "similar" are two different brands or how similar are two different attributes. The last can be useful because it can be helpful while defining and analyzing the set of attributes assigned to a brand.

They also propose the study of the brand perception using the following methodology: given a sentence with brand and masked attribute word, they use a pretrained LM (with Masked Language Model[2] head) to predict words at the masked position. If a model predicts the expected[3] attribute in the top-5 position, then it can be inferred that the model representations have captured the corresponding affect association. Another interesting experiment they realize to further analyze sensitivity to context, is that they perturb the sentences introducing nonsensical words. For example:

"I should play Nintendo because it is [MASK] ."
"I snap play Nintendo ya it is [MASK] ."

In order to do all these experiments they use five different models: BERT, RoBERTa, DistilBERT, ALBERT, and BART. Both synthetic and real-world data are employed to evaluate how these models understand and encode associations between brands and various attributes. The study finds consistent associations between brands and attributes across models, although these do not always align with consumer or intended brand perceptions. The results also demonstrate varying degrees of effectiveness, with some models better capturing nuances in brand perception and that perturbations in sentences moderately influence these associations.

In Scherrer et al., 2024 they investigate how LLMs encode and process moral beliefs. The study comprises two main components: the development of a statistical method to elicit encoded beliefs and the application of this method to a large-scale survey of moral scenarios.

[2] A Masked Language Model (MLM) is a type of language model used in natural language processing that is trained to predict missing or "masked" words in a sentence.

[3] They refer to surveys conducted by Young and Rubicam, a renowned global marketing and communications company that has grown to become one of the largest advertising agencies in the world.

About the survey design, the researchers designed a survey with 1,367 moral scenarios, including both high-ambiguity (e.g., "Should I tell a white lie?") and low-ambiguity (e.g., "Should I stop for a pedestrian on the road?") cases. Given a scenario, they create six different question forms: choosing between an A/B scenario, repeating the scenario that has been chosen or answering yes or no when asked about the preference from one option to the other. Not only they formulate the same scenario in three different ways, but also they randomize the order in which the options appear to the LLM.

They define two metrics regarding the consistency and entropy of the LLM. Note that they ask the LLM to answer the same questions multiple times, say M. For consistency, given a specific question format, they calculate the likelihood of each option across the M times they ask the question, counting the times it responds with one option or another and dividing by M, they call this Action Likelihood (AL). If they do this for each different way of asking the question and then average for each different type of question, we obtain what they call Marginal Action Likelihood (MAL). In other words, we are obtaining the probability that the LLM will respond with a particular option regardless of the question format used. Finally, to calculate the first metric, they compare the difference between the average probability of each option (MAL) with the probability obtained in the initial calculation, where for each question format a probability was obtained. They compute this difference between probability distributions using the concept of divergence.

Regarding the entropy metric, once the AL is calculated, they compute the entropy for each question. In other words, for each question (of a specific format), they calculate the entropy per question. This is called Action Entropy (AE). Similarly to before, they compute the Marginal Action Entropy (MAE), which is nothing more than calculating entropy, but now using the probabilities obtained in MAL. That is, entropy is obtained for each scenario independently of the different ways they ask the same question, since averaging has already been performed beforehand. This will be the entropy metric they will ultimately use to compare different models.

The survey was administered to 28 different LLMs, both open and closed-source. Analzying the previous metrics they conclude that in scenarios with clear moral implications, most LLMs chose actions aligning with commonsense moral principles. However, in ambiguous scenarios, the models often showed uncertainty, indicating the complexity of encoding nuanced moral judgments.The study found that the responses of some models were highly sensitive to how questions were worded, leading to inconsistencies in moral decision-making. Closed-source models tended to show more agreement with each other in ambiguous scenarios compared to open-source models, suggesting differences in how moral beliefs are encoded across different model architectures.

Both studies underscore the significant capabilities of LLMs in analyzing complex human attributes such as brand perception and moral beliefs. The research on brand perception illustrates the practical applications of LLMs in marketing and the importance of addressing model biases. Meanwhile, the study on moral beliefs emphasizes the nuanced understanding required for ethical AI applications and the variability in model responses depending on scenario ambiguity and question phrasing. Together, these studies provide a comprehensive overview of the state of

the art in leveraging LLMs for understanding human-centric attributes and decision-making processes.

Chapter 3

Methodology

Despite considering the papers mentioned earlier in the introduction, it is important to note that we have not strictly followed the metrics explained in them. This decision was due to some cases where the terminology used was somewhat confusing and certain decisions were not fully justified. Nevertheless, the methodology described below is inspired by these papers and incorporates knowledge in psychometric analysis and brand perception.

Before describing our experimental procedures, it is crucial to understand the methodology followed. The process can be divided into three main sections:

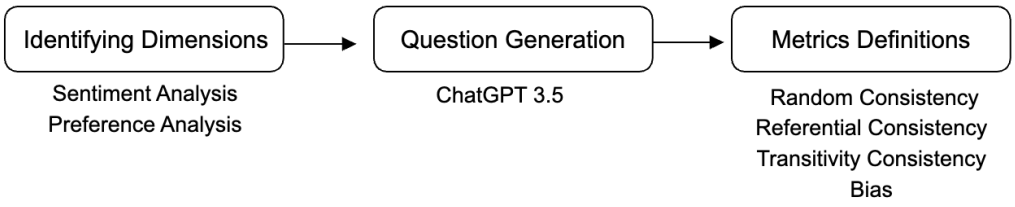


FIGURE 3.1: Outline of the methodology followed in this study.

First, we define the dimensions we want to measure. Next, we generate the questions. Finally, we define the metrics that will help us measure the described dimensions using the questions. Let us explain in depth each section.

3.1 Identifying Dimensions

In our case, we can differentiate between two distinct dimensions: sentiment and preference. Let us start by explaining the first dimension.

By sentiment, we mean: given any brand, what idea/position does the LLM have about it? There are different ways to approach this question:

1. *Random Consistency*. Does the LLM respond the same if I ask exactly the same question?

2. *Referential Consistency*. Does it respond the same regardless of the question modality?

3. *Bias*. Does it tend to value my brand positively? Or does it do so negatively?

Taking into account these three initial questions, we will define the three sentiment metrics later on. Regarding the second dimension, when we talk about preference, we mean: given two brands, which one does the LLM prefer? Once we know which brand does the LLM prefer we can evaluate whether the LLM is "logical." In other words, we assess if it satisfies the property of transitivity. For example, if it tells us that B_1 is better than B_2 and that B_2 is better than B_3, it should logically follow that B_1 is better than B_3. To verify this property, which we will call **Transitivity Consistency**, we will ask the LLM about three different brands. By examining the responses for all pairwise comparisons among the three brands, we can determine if the LLM maintains logical consistency in its preferences. If it does, this would confirm that the LLM's decision-making process is transitive and, therefore, logically sound.

3.2 Question Generation

According to what we explained about the creation of a test, it makes sense that to get an idea of what perception an LLM has about any brand, we ask questions that reference to these characteristics (see Section 2.2). For instance:

1. Do you consider that the brand B_1 offers a satisfactory shopping experience?
2. Do you think that the brand B_1 is viewed as a trustworthy brand by consumers?
3. Have you had good experiences with products/services from the brand B_1?
4. Do you think that the brand B_1 has a distinctive personality?

Another crucial consideration for test designers is how to formulate questions effectively. There are various methods to inquire about the same concept: dichotomous questions, open-ended questions, multiple-choice questions, Likert-type scales, and more. In this study, for the sake of simplifying response analysis, dichotomous questions and Likert-type scales[1] have been selected. However, it is important to note that LLMs are sensitive to the phrasing of questions. Similar to the approach in Scherrer et al., 2024, we will employ a variation of dichotomous questions. Instead of simply asking for a Yes/No response, we will reframe these questions to prompt the LLM to choose between options A) or B), where A corresponds to "Yes" and B corresponds to "No". Additionally, we will modify the Likert scale questions to range from 1 to 4.

In summary, we will first develop a set of 100 questions that address the aforementioned characteristics of a brand. Subsequently, we will adapt these questions in the following ways:

1. Yes/no questions.
2. A/B questions.
3. Likert with 4 points.

[1]These questions present a statement and ask respondents to indicate how much they agree or disagree on a scale, usually 5-point (for example, from "Strongly Disagree" to "Strongly Agree").

4. Likert with 5 points.

5. A vs B questions.

Let us examine an example of a question and how it would change in relation to the previous list. Let B_1, B_2 represent two brands, and a_j denote different options. The corresponding variations of the same question are:

Example 3.1.2

1. Have you heard of the brand B_1 before? a_1 : Yes a_2 : No

2. Have you heard of the brand B_1 before? a_1 : A) Yes a_2 : B) No

3. I have heard of the brand B_1 before. $a_1 : 1, a_2 : 2, a_3 : 3, a_4 : 4, a_5 : 5$

4. I have heard of the brand B_1 before. $a_1 : 1, a_2 : 2, a_3 : 3, a_4 : 4$

5. Which brand are you more familiar with: B_1 or B_2? $a_1 : B_1$ $a_2 : B_2$

Where in questions 3 (Likert 4) and 4 (Likert 5) we do the following association:

Option	Likert 5	Likert 4
Strongly disagree	1	1
Disagree	2	2
Neither agree nor disagree	3	-
Agree	4	3
Strongly agree	5	4

TABLE 3.1: Likert Scale Option Mappings.

Additionaly, the fifth question of the previous example correspond to the preference dimension. For this dimension, we will use the same question bank as before, but instead of asking about a single brand, we will include both brands in the questions. In order to do this we will also question the LLM in four different ways, for instance, the modifications to the 5^{th} question of the example would be:

1. Are you more familiar with B_1 than with B_2. a_1 : Yes a_2 : No

2. Are you more familiar with B_1 than with B_2. a_1 : A) Yes a_2 : B) No

3. I am more familiar with B_1 than with B_2. $a_1 : 1, a_2 : 2, a_3 : 3, a_4 : 4, a_5 : 5$

4. I am more familiar with B_1 than with B_2. $a_1 : 1, a_2 : 2, a_3 : 3, a_4 : 4$

We also incorporate a variation in the order of questions to avoid potential biases. Specifically, if we ask the same question m times, we will ask $\frac{m}{2}$ times as B_1 vs B_2 and the other half as B_2 vs B_1.

For the question generation process, we utilized ChatGPT 3.5 to generate a set of questions related to brand analysis. Initially, we requested the model to provide 100 dichotomous questions, with careful consideration given to the discussed characteristics of unidimensionality. After that, we conducted an analysis of these questions, identifying similarities and generating additional questions to finalize a set of 100

questions. Once the dichotomous questions were prepared, the next step involved slight modifications (following the framework of the Example in 3.2) to adjust both the questions and their corresponding answers.

It is important to mention that all questions have been generated with a positive connotation. In other words, for any question, if the answer is Yes, it is understood to be "good" for the brand. Let us provide an example to clarify this important aspect:

Suppose the question is:

Do you think that the brand B_1 has a good brand image?

If the response is Yes (Yes, A) or (4,5) or (3,4), this denotes a positive perception of the brand B_1. It should be noted that the question could also be formulated in the opposite manner:

Do you think that the brand B_1 does not have a good brand image?

Indeed, despite not incorporating this variant in our experiments, we consider it to be a very good strategy. We, therefore, encourage future studies to include it and even vary questions with both positive and negative connotations.

3.3 Metrics Definitions

Once we know the typology of our questions, we define how we aim to measure the dimensions we explained previously to study the brand perception of an LLM in relation to a brand.

Figure 3.2 illustrates the general framework of the sentiment dimension we aim to evaluate. As we elaborate on how to calculate the metric for each initial question, this graph will become clearer and help in understanding the different metrics.

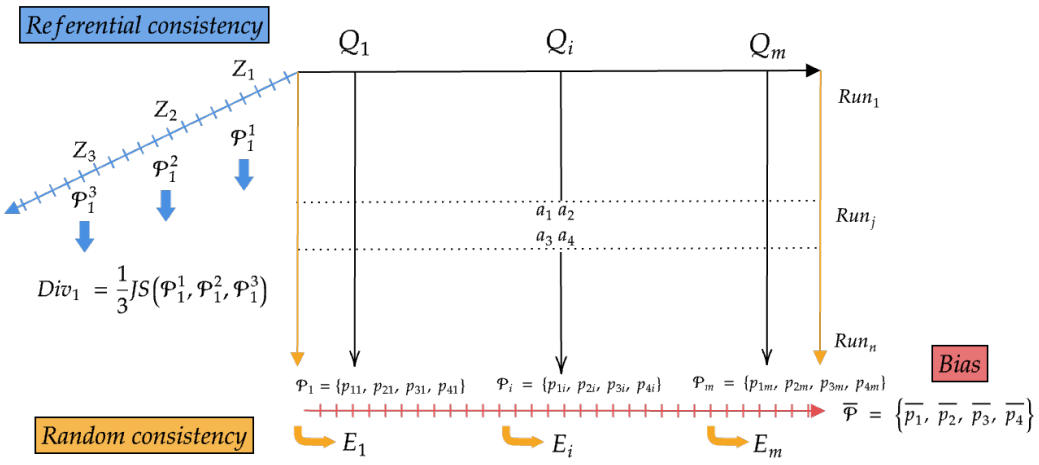


FIGURE 3.2: General framework for sentiment dimension metrics. The idea is that we have this scheme for each different question format. Q_i represent question i, E_i represents entropy for question i and Z_k represent different question formats.

3.3.1 Random Consistency

Suppose a question is posed repeatedly to an LLM in the same format. In this context, we define **random consistency** as the LLM's ability to provide the same answer to the question each time it is asked. Note that in this metric, we will focus on a single question format.

Mathematically let $\mathcal{Q} = \{q_1, q_2, \ldots, q_m\}$ denote a set of m questions and $\mathcal{A} = \{a_1, a_2, \ldots, a_l\}$ be a set of l[2] different answers, whose options are available for each question. Each question $q_i \in \mathcal{Q}$ is posed n times. Let A_{ik} denote the answer provided by the language model $\mathcal{L}$ to the k-th instance of the i-th question, where $A_{ik} \in \mathcal{A}$.

Hence, let us explain how we measure the random consistency of $\mathcal{L}$ given a question q_i. Note that this question is posed to the language model n times, so we have K answers $\{A_{i1}, A_{i2}, \ldots, A_{in}\}$. The frequency of each unique response to the question q_i and its respective probability[3] can can be calculated as

$$\begin{gathered} f_j = \sum_{k=1}^{n} \mathbb{1}(A_{ik} = a_j) \quad j \in \{1, 2, \ldots, l\} \\ p_j = \frac{f_j}{n} \end{gathered} \tag{3.1}$$

Now that we have the probability of each possible answer we would like to measure how random these answers are. The way we propose it is through the entropy. Entropy is a concept from information theory that quantifies the uncertainty or unpredictability of a random variable. In the context of a language model's responses, entropy can be used to quantify the variability in the answers given to a repeated question.

Given a discrete random variable X which takes values in the set $\mathcal{X}$ and is distributed according to $p\colon \mathcal{X} \to [0, 1]$, the entropy is

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log(p(x))$$

Having into account the scenario described above, note that as our questions will have 2 (dichotomous) or 4, 5 (Likert) possible options, our random variable (q_i) will follow a Binomial or a Multionmial distribution, respectively. Therefore, the final entropy per question will be

$$H(q_i) = -\sum_{j=1}^{l} p_j \log(p_j), \quad i \in \{1, 2, \ldots, m\} \tag{3.2}$$

We find the maximum entropy when all events are equiprobable, that is, when $p_j = \frac{1}{l}\ \forall j \in \{1, 2, \ldots l\}$. In such case, the maximum entropy is $-\log(\frac{1}{l})$ (see Figure 3.3). Conversely, we find the minimum entropy when an event is clearly preferable among the others; for example, in the case of a dichotomous variable when one event has probability 1 and the other 0. Note that $log(0)$ it is not defined, hence in our experiment in order to make it work we add a little tolerance.

[2] If we change the format question the number of available answer per questions can be different, for instance, Likert versus dichotomous questions.

[3] We estimate the probability by computing the frequentist probability of each option. It is intuitive that we would want n to be as large as possible. Nevertheless, we are bound by computational constraints.

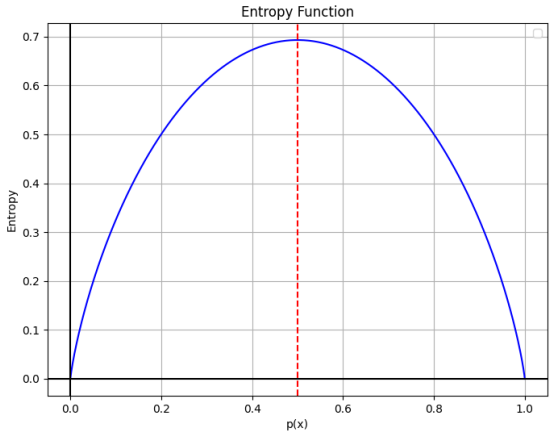


FIGURE 3.3: Entropy for a dichotomous variable. It reaches its maximum when the probability of one event is $\frac{1}{2}$.

The interpretation of the entropy is as follows: if the entropy $H(q_i)$ is high, the model's answers are highly variable, indicating low random consistency. The model is less consistent because its responses are spread out across different possible answers. Whereas if the entropy is low (close to 0), the model's answers are highly consistent, indicating high random consistency. The model tends to give the same answer most of the time.

Entropy allows us to measure how randomly consistent an LLM is given a specific question and question format. Initially, we expect an LLM to be consistent in its responses, so we would anticipate low entropy for each question. However, in terms of brand analysis, high entropy for each question could imply that the LLM does not have a clear position or preference regarding the brand in question. On the other hand, very low entropy would suggest that the LLM has a clear and consistent stance on the brand. The interpretation of whether this stance is positive or negative would depend on the specific results.

3.3.2 Referential Consistency

Suppose a question is posed to an LLM with n different question formats (as seen in 3.2). In this context, we define **referential consistency** as the LLM's ability to respond to the question congruently independent of the question and answer format.

Mathematically, consider n different formats of asking the same question. So, for a certain question q_i we will have n different variations of the same question, say $\{q_{i1}, q_{i2}, \ldots, q_{in}\}$. Returning to the definition we gave of the probability of an option in a certain question 3.1, we can now calculate $n(l_1 + l_2 + \ldots + l_n)$ different probabilities, each depending on how the question is formulated (different $\mathcal{Q}$) and taking into account that the set of answers ($\mathcal{A}$) will also be different. Thus, we can obtain a vector of probabilities $P_i^j = \{p_1, p_2, \ldots, p_{l_j}\}$ for question q_i and format j (see Figure 3.2).

To clarify further, let us consider a scenario where questions are formulated in different formats but maintain the same criteria for answers, or at the very least, the

same number of possible answers.

Let us introduce the Kullback-Leibler (KL) divergence, which measures how much a probability distribution differs from another. It quantifies the amount of information lost when using one distribution to approximate the other one. The lower the values, the more similar both distributions are (conversely, greater values indicate greater distribution differences).

Formally, for two discrete probability distributions P_1 and P_2 defined on the same probability space, the KL divergence from P_2 to P_1 is given by:

$$D_{\mathrm{KL}}(P_1 \parallel P_2) = \sum_{x \in \mathcal{X}} P_1(x) \log \frac{P_1(x)}{P_2(x)}$$

where:

- $\mathcal{X}$ is the set of possible outcomes.
- $P_1(x)$ is the probability of outcome x under distribution P_1.
- $P_2(x)$ is the probability of outcome x under distribution P_2.

However, the KL divergence is not symmetric[4], so we introduce the Jensen-Shannon divergence (JSD) defined as:

$$D_{\mathrm{JS}}(P_1 \parallel P_2) = \frac{1}{2} D_{\mathrm{KL}}\left(P_1 \parallel M\right) + \frac{1}{2} D_{\mathrm{KL}}\left(P_2 \parallel M\right)$$

where M is the average of the two distributions P_1 and P_2:

$$M = \frac{1}{2}(P_1 + P_2)$$

Another advantage of the JS divergence is that it is bounded between 0 and 1 when using log base 2, while the KL divergence is not. Same as with the KL divergence, a smaller value indicates more similarity, while a larger value indicates more dissimilarity, obtaining that $D_{\mathrm{JS}}(P_1 \parallel P_2) = 0$ if and only if $P_1 = P_2$.

Therefore, considering that we have the probability of each option for each type of question, we can compare how different these probabilities are from each other in the following way:

Let $\{P_i^j\}_{j=1}^k$ be a set of k discrete probability distributions defined on the same probability space $\mathcal{X}$. In our case, for each question q_i, P_i^j represents the vector of probabilities using one format and P_i^s using another format. Note that $\mathcal{X}$ is the same for all P_i^j because we are assuming that they have the same number of possible answers[5]. The total JS divergence between these distributions is:

$$D_{q_i} = \frac{2}{k(k-1)} \sum_{1 \le j < s \le k} D_{\mathrm{JS}}(P_i^j \parallel P_i^s) \quad i \in \{1, 2, \ldots, m\} \tag{3.3}$$

Previously, we only considered cases where we have the same number of options to respond to a question. But what happens if this is not the case? Earth Mover's

[4]The fact that KL divergence is not symmetric is not an inconvenience per se, but the reader will see how using a symmetric divergence can simplify calculations in the Experimental section.

[5]This can be easily achieved, for example, by dichotomizing Likert variables.

Distance (EMD) provides a solution. This allows us, for example, to compare the difference between a Likert scale and a dichotomous scale. However, normalization is necessary to ensure a fair comparison.

Let us explain a little bit (Hulet, 2024) the formulation behind this concept. A general idea is the following: suppose we have two distributions, P and Q, and we want to know how different they are by transforming P into Q and measuring how much total work was done. In other words, the number of units we have moved times the distance moved to make the transformation possible. There are many different ways to do that, but we want to obtain the minimum amount of work required for the most efficient transport plan[6].

We can reformulate this problem as an optimization problem as follows:

$$\begin{aligned} &\min_X \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} \cdot x_{ij} \\ \text{subject to} \quad &\sum_{j=1}^{n} x_{ij} = p_i, \quad 1 \le i \le m \\ &\sum_{i=1}^{m} x_{ij} = q_j, \quad 1 \le j \le n \\ &x_{ij} \ge 0, \quad 1 \le i \le m, 1 \le j \le n \end{aligned}$$

Where we can think x_{ij} as a matrix that has distribution P values as columns and distribution Q values as rows. This matrix will have a value for every combination of observations between the two distributions, assigning a zero if any combination between two observations has been done (indeed, it is a possible transport plan). About d_{ij}, it is also another matrix, usually called cost matrix, where the distance between each observation from the first distribution to the second one is stored. And p_i and q_j represent the values in each distribution, P and Q, respectively. Hence, now the objective function makes sense, as we want to obtain the optimal transport plan X that minimizes the dot product between the transport plan and the cost matrix.

The constraints are quite intuitive: the first one guarantees that every observation from the first distribution is moved to the second distribution precisely once, while the second constraint ensures that the resultant transformed distribution matches the second distribution. The last constraint ensures that the elements of the matrix transport plan are non-negative, as it is not possible to transport negative quantities.

Therefore, in the end, this distance helps us measure the difference between two distributions as we were discussing earlier, but with the advantage that we can compare with different probability distributions.

With both approaches, we are measuring how much the different ways of asking the LLM vary for each question. A desirable outcome would be obtaining a low divergence (D_{q_i}) for each question. This would indicate coherence regardless of

[6]Here we call transport plan to the set of moves we make to transform one distribution into the other.

the response format. Alternatively, a high total divergence per question would suggest that the LLM provides different responses when the question/response format changes. Another approach is to assess which question format poses the most difficulty for the LLM by summing only the total Jensen divergences where one question format interferes and excluding the other ones (this is seen in Figure 4.4).

3.3.3 Bias

Suppose m questions are posed to an LLM n times. In this context, we define **bias** as the LLM's tendency to answer the questions affirmatively or negatively through the n times we ask.

We return to Formula 3.1. We will now move horizontally (see Figure 3.2) . That is, we want to calculate the average degree to which the LLM is in favor or not of the brand. To do this we will simply average the different p_j. Let us denote p_{ji} and f_{ji} as the p_j and f_j obtained in question q_i for option j, where $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, l\}$.

$$\begin{gathered} \overline{p}_j = \frac{1}{m} \sum_{i=1}^{m} p_{ji} \forall j \in \{1, 2, \ldots, l\} \\ \overline{f}_j = \sum_{i=1}^{m} f_{ji} \forall j \in \{1, 2, \ldots, l\} \\ Preference = a_i \text{ such that } \max_i \{\overline{p}_1, \overline{p}_2, \ldots, \overline{p}_l\} \\ Mode = a_i \text{ such that } \max_i \{\overline{f}_1, \overline{f}_2, \ldots, \overline{f}l\} \end{gathered} \tag{3.4}$$

Let us do an example, in the first Table we have $\mathcal{Q}_1 = \{q_1, q_2, q_3\}$ and $\mathcal{A}_1 = \{Yes, No\}$ whereas in the second Table (right one) we have $\mathcal{Q}_2 = \{q'_1, q'_2, q'_3\}$ and $\mathcal{A}_2 = \{1, 2, 3, 4\}$

Run/Question	q_1	q_2	q_3
1	No	Yes	Yes
2	Yes	Yes	No
3	No	Yes	Yes

Run/Question	q'_1	q'_2	q'_3
1	1	3	2
2	4	3	2
3	1	4	1

TABLE 3.2: Potential use case with two different question formats.

$$p_{11} = \frac{1}{3}, \quad p_{21} = \frac{2}{3}, \quad p_{12} = \frac{3}{3}, \quad p_{22} = \frac{0}{3}, \quad p_{13} = \frac{2}{3}, \quad p_{23} = \frac{1}{3}$$

$$\overline{p}_1 = \frac{2}{3}, \quad \overline{p}_2 = \frac{1}{3}, \quad \overline{f}_1 = 6, \quad \overline{f}_2 = 3$$

And in the second one

$$\begin{aligned} p_{11} &= \frac{2}{3}, & p_{21} &= \frac{0}{3}, & p_{31} &= \frac{0}{3}, & p_{41} &= \frac{1}{3}, \\ p_{12} &= \frac{0}{3}, & p_{22} &= \frac{0}{3}, & p_{32} &= \frac{2}{3}, & p_{42} &= \frac{1}{3}, \\ p_{13} &= \frac{1}{3}, & p_{23} &= \frac{2}{3}, & p_{33} &= \frac{0}{3}, & p_{43} &= \frac{0}{3} \\ \overline{p}_1 &= \frac{1}{3}, & \overline{p}_2 &= \frac{2}{9}, & \overline{p}_3 &= \frac{2}{9}, & \overline{p}_4 &= \frac{2}{9} \\ \overline{f}_1 &= 3, & \overline{f}_2 &= 2, & \overline{f}_3 &= 2, & \overline{f}_4 &= 2 \end{aligned}$$

In the first table we obtain a preference for option 1, that is, "Yes" ; whereas in the second table we obtain a preference for option 1, that is, "1", which can be thought as *strongly disagree*.

We will compute this for each question format. Hence, we will obtain a preference for one option per for different question format. Here, we will be able to compare how the preference varies depending on the question format used and also, and more interesting, if we do the average betweeen all the possible formats[7] we can conclude the average preference for a brand.

Note that if we follow the same scheme for more than one brand, we can compare the level of preferences between brands.

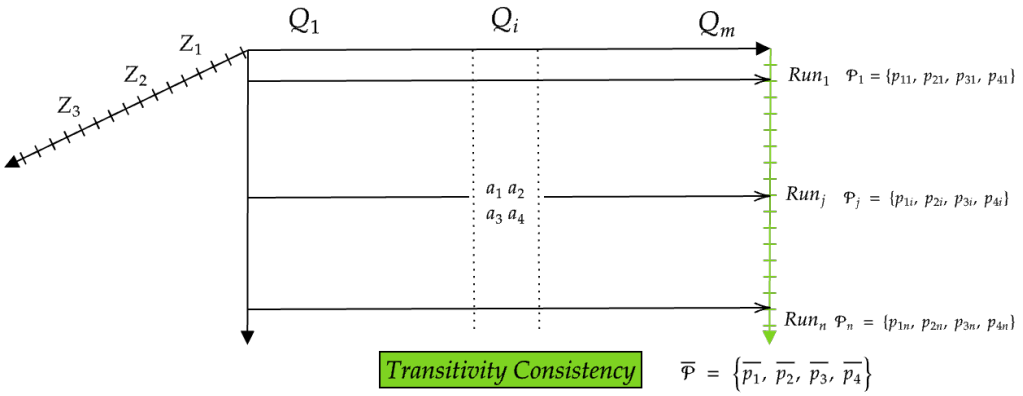


FIGURE 3.4: General framework for preference dimension metric. Again, we have this scheme for each different question format Z_k.

3.3.4 Transitivity Consistency

This last subsection regards to the preference dimension, which involves directly asking the LLM its preference for one brand over the other. We will proceed in a similar way we did with the *Bias* metric. The similarity lies in that we will also move horizontally (see Figure 3.4), but this time it will be the first movement we make, unlike the first diagram (see Figure 3.2). In other words, for each test we conduct, we will calculate the probability of each option for each question (horizontal) and then group vertically by averaging over the number of times we repeat the same test.

Mathematically, suppose m questions are posed to an LLM n times, within a same question format. We will now start measuring in an horizontal way. That is, for each time we ask the m questions, we will estimate the probability, in the same way we did in 3.1 but for each run. That is, following the same notation as in 3.1 for each run $s \in \{1, \ldots, n\}$:

$$\begin{gathered} f_{sj} = \sum_{k=1}^{m} \mathbb{1}(A_{ik} = a_j) \quad j \in \{1, 2, \ldots, l\} \\ p_{sj} = \frac{f_j}{m} \end{gathered} \tag{3.5}$$

[7] In order to do this we first need to establish a relation between options among different question formats. See 4.3.

Therefore, we will obtain a set of probabilities for each option. If we now compute the average for each option among all the different runs we obtain a final probability for each option for all the n runs[8]:

$$\overline{p_j} = \frac{1}{n} \sum_{s=1}^{n} p_{sj} \quad j \in \{1, 2, \ldots, l\} \tag{3.6}$$

Note that we can repeat this procedure for different question formats. If we also combine this by considering the other formats, i.e., by averaging the different question formats, we will obtain an overall average preference for the brand.

Let us discuss the example when we only have two options, for every different question format. Additionally, let us also suppose that we can categorize the two options as "Yes" and "No". In this context, if we consider the average of "Yes", we obtain the average probability with which the LLM prefers brand B_1 over brand B_2.

This value is what we will use to determine whether the LLM prefers brand B_1 or brand B_2, and it will allow us to subsequently calculate the Transitivity Consistency.

[8]It is worth mentioning that in this case, out of the n runs, half of them ask B_1 versus B_2 and the other half B_2 versus B_1. When quantifying the responses, this variation must be taken into account

Chapter 4

Experimental setup

4.1 Groundwork

We will start explaining the manner in which we pose questions to the LLM. From the previous section, we already know that we ask 100 questions for each different type of test. However, until now, we have not explained how we present the set of questions from a test to the LLM. There are two options: asking questions sequentially one after the other or asking all the questions at once.

We understand that perhaps the most effective way to ask the questions is sequentially, as this way the LLM does not consider previous questions nor is it influenced by other questions when responding to a particular question. However, if a test is given to a person, the questions would be presented all at once, influencing the decision of one to the response of another question.

It is worth noting that initially, we tried both approaches, but we ultimately opted for the second option due to time constraints: if we asked the test questions sequentially, it took much longer to obtain the 100 responses for a single test. Considering that we have 4 different question formats, that we will study more than one brand, and that we will repeat the same test more than once, this results in a considerable increase in time.

Type of Question	**Prompt's head**
Yes/No	*"""I'm doing a brand perception study. Can you answer the following questions with yes or no? :* *1. Have you heard of the brand B_1 before?* *2. Have you used products/services from the brand B_1 in the last year?* *3. Do you believe that the brand B_1 is nationally recognized?* *... """*

TABLE 4.1: Prompt's head in Yes/No questions.

Thus, having decided to ask the questions at once, we only need to select which models we want to use. However, here we encounter another limitation: *context window*. The context window of an LLM refers to the maximum amount of text (in terms of tokens) that the model can process and consider at one time. For instance, the context window of Lama 2 is 4096 tokens, which is very small for our tests. Taking all this into account, we ended up selecting the following models for our tests: ChatGPT-3.5, ChatGPT-4, and Gemma. The selection of ChatGPT-4 is partly due to

curiosity about the difference between the continuity of models. Let us see a brief summary of these models:

# Parameters	Acces	Provider	Context Window	Model
2B	Open Source	Google	8,192	*gemma-2b-it*
Unknown	API	OpenAI	16,385	*gpt-3.5-turbo*
Unknown	API	OpenAI	128,000	*gpt-4o*

TABLE 4.2: General information about used models.

When prompting LLMs, there are various parameters that can be adjusted. This has the potential to significantly alter how the LLM responds. Let us briefly discuss some of the parameters we have considered (Walia, 2023):

1. **temperature**: It influences the randomness of the generated responses. Ranging from from 0 to 1, the higher the value the more diverse the responses. Conversely, the lower the value, the more focused and deterministic the responses become.

2. **max_tokens**: It allows you to control the length of the generated response by specifying the maximum number of tokens the model will generate. For example, if max_tokens is set to 50, the model will generate up to 50 tokens in its response. If the complete response exceeds this limit, it will be truncated at 50 tokens.

3. **top_p**: It controls the diversity and quality of the responses by limiting the cumulative probability of the most likely tokens. Ranging from 0 to 1, the higher the value the greater variety of tokens is allowed, resulting in more diverse responses. While lower values yield more constrained answers.

To attempt a more balanced comparison, we have maintained the same parameter values across all the LLMs, setting: temperature to 1, max_tokens to 1000 and top_p to 0.95.

Another interesting aspect to comment on is the analysis of responses obtained from different LLMs. As expected, there are many responses that are not valid for our study, which implies repeating the questions until the LLM responds according to our criteria. An example of output that we do not accept during the experiments is:

1. As an AI developed by OpenAI, I don't have personal experiences or opinions.

2. Specially with Gemma, we found out that when asked for a Likert 4 type scale answer, the LLM answered many times giving a 5 (which was not an available option).

3. Providinig a general answer (Yes or No) instead of answering to each of the questions.

An interesting analysis would be to compare which LLM tends to provide responses that we consider incorrect, taking into account the question format as well

as the brand.

Let us now explain the procedure we followed to record the responses of the LLM to the different tests. As mentioned earlier, there are different types of tests (see example questions in 3.2). To analyze the responses, we have considered the following classification:

Type of Question	Options	Classification
Likert 5	1,2,3	0
	3,4,5	1
Likert 4	1,2	0
	3,4	1
A/B	B	0
	A	1
Yes/No	No	0
	Yes	1

TABLE 4.3: Classification of LLM's responses.

As can be observed, in the Likert scale 5 option, option 3 appears twice, classified as both 0 and 1. This is because option 3 is: Neither agree nor disagree. Therefore, each time the LLM returned a 3, we assigned it (with a 0.5 probability) as either a 0 or a 1. It is important to note how this matches the definition we provided of positive connotation in 3.2.

In this way, the analysis is greatly simplified since for each response to a question we only end up storing a 0 or a 1. This information can then be easily stored in matrices for further analysis. Nevertheless, we are also losing some information (especially with Likert scale questions) although if we want to retain these Likert scale responses and not perform the conversion, the EMD alternative can be used instead of the Jensen divergence (as explained in 3.3.2).

The brands chosen for this study are *Apple, Samsung*, and *Huawei*. We will analyze the metrics described earlier comparing Apple to Samsung for the sentiment dimension[1] and a three-way comparison between Apple, Samsung, and Huawei for the preference dimension.

4.2 Results

In this section, we will discuss the results obtained from applying the aforementioned metrics for the different selected brands and across different LLMs. At the end of this section, we will summarize the results obtained, categorizing ultimately the perspective that the LLM has on a brand. We remind the reader that each question format contains 100 questions and that we repeat the same exactly question format 30 times.

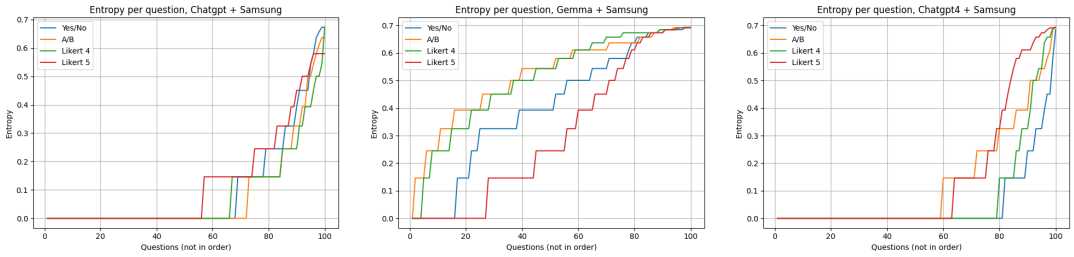
[1] In fact, we have performed all the calculations for Huawei as well. However, to better explain the results, we decided to discuss the results for this dimension using only two brands. The final results section will present the findings for Huawei.

4.2.1 Random Consistency

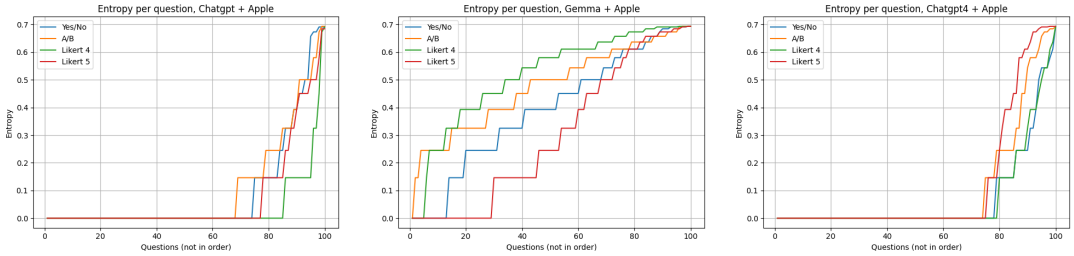
As we are interested in categorizing the LLM's opinion regarding brands, the final result we aim to obtain from this metric is a general measure of how much entropy, overall, an LLM exhibits when asked about a specific brand. However, leading up to this conclusion, there is a set of alternative questions that are interesting and provide us with information about the performance of the LLM. An example of these questions is:

1. Which questions exhibit more entropy? Does this apply uniformly across all different format questions?
2. Which question format has the most entropy?
3. Which LLM exhibits more entropy?

Through the analysis of each graph we present, we will address each of these questions. To start with, let us study how entropy is distributed among the questions. In the following graph (Figure 4.1), we order the questions based on entropy, from lowest to highest.



(A) ChatGPT3.5 & Samsung (B) Gemma & Samsung (C) ChatGPT4 & Samsung



(D) ChatGPT3.5 & Apple (E) Gemma & Apple (F) ChatGPT4 & Apple

FIGURE 4.1: Entropy Comparison By Question. Comparison among different models and brands, ordering questions from lowest to highest entropy.

Regarding ChatGPT3.5 with questions about Samsung, we observe that more than half of the questions (across all four different formats) exhibit nearly zero entropy. This indicates that in all these questions, the model is very certain about which option to choose. As for ChatGPT with Apple, it is even more certain, as there are more questions with zero entropy, reaching over 80 for the Likert 4 format. In general, in both cases, a low total entropy is demonstrated.

In the case of Gemma, the initial impression is of significantly higher total entropy compared to ChatGPT3, regardless of the question format used. We notice that

in none of the cases (neither Apple nor Samsung) are there more than 30 questions with nearly zero entropy (for any of the four formats), which leads to a considerably higher entropy than that obtained with ChatGPT3.

Finally, ChatGPT4 exhibits a behavior very similar to that of ChatGPT3. There are many questions with low entropy for both Apple and Samsung, and around 30 to 25 questions that contribute to the higher entropy for ChatGPT4. Overall, it has lower entropy compared to Gemma.

Looking at the previous plot, a natural question arises: which questions are causing the higher entropy? To determine which questions have the highest and lowest entropy, we have summed the entropy obtained in each different format for each question. Let us see some examples:

1. **Gemma**: 11. Do you think that the brand Samsung is accessible in terms of price?
2. **ChatGPT3** : 18. Have you had any interaction with the brand Apple on social media?[2]
3. **ChatGPT4** : 43. Do you think that the brand Apple is accessible to different demographic groups?

On the other hand, the questions with least entropy are:

1. **Gemma**: 100. Do you consider that the brand B_1 is seen as a brand that cares about its customers?[3]
2. **ChatGPTs**: 1. Have you heard of the brand B_1 before?[4]

Here, we present just one example of the questions for a possible analysis. Note that this analysis could also help us discard or improve our question bank. In fact, an interesting study would be to categorize which questions have the most/least entropy according to the classification mentioned in Section 2.2.

In Figure 4.2 , we have summed the entropy of each question by question format and by brand. It is evident that Gemma exhibits a much higher total entropy compared to ChatGPTs. This plot also allows us to identify which question formats exhibit more and less entropy. If we look at the last bars for each format question, we can see how for both Apple and Samsung the A/B format exhibits the most entropy while the format with lower entropy for Samsung is Yes/no and for Apple it is Likert 5. It is curious how the A/B format exhibits the highest entropy, while the Yes/No format, being relatively similar, shows the lowest entropy.

Furthermore, the stars in the bars allow us to see in a practical way the vertical comparison between Apple and Samsung within the same LLM. In other words, the bar where the star is located indicates which brand has higher entropy. Initially, we can see that in the majority of cases Samsung (which is always at the top) exhibits more entropy than Apple. To establish a better comparison, let us examine the overall results without distinguishing question format:

[2]This questions is also the one with the highest entropy for Samsung.
[3]This question is the one with least entropy for all brands,even Huawei.
[4]It applies for both Samsung and Apple.

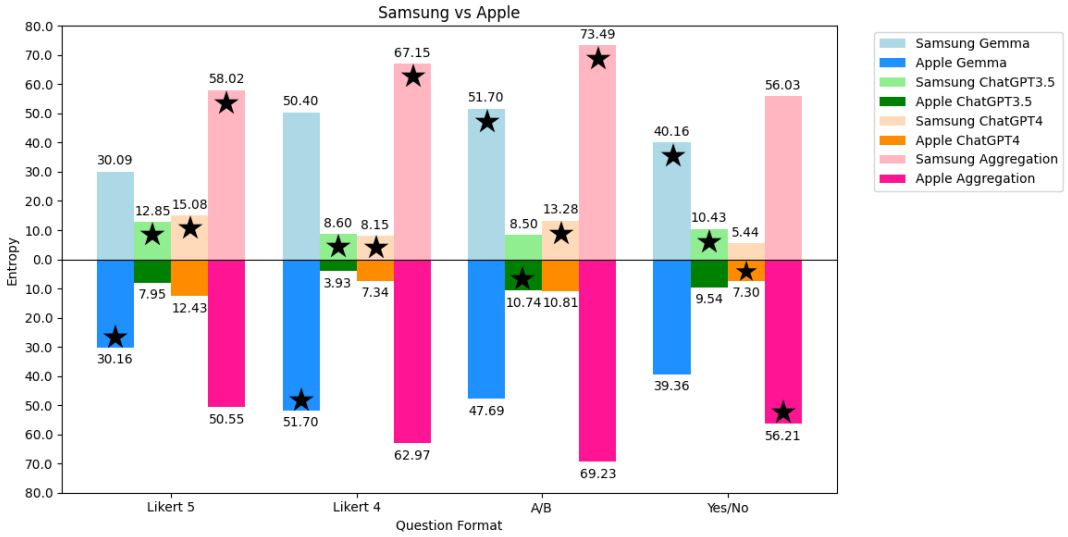


FIGURE 4.2: Entropy Comparison By Question Format. Accumulation of entropy by model, brand, and question format. The top part of the figure refers to Samsung, and the bottom part to Apple. The pink bars show the accumulation of entropy for each LLM by question format. The ★ indicates which entity has more entropy in the vertical comparison between Samsung and Apple.

Brand/LLM	Gemma	ChatGPT3.5	ChatGPT4	Total
Apple	168.9	32.2	37.9	239
Samsung	172.4	40.4	42	254.8
Huawei	172.8	157	113.3	443.1
Total	514.1	229.6	193.2	936.9

TABLE 4.4: Random consistency by brand and LLM results.

The table clearly shows that Samsung exhibits slightly higher entropy than Apple across all LLMs, and that the difference between Huawei and the two brands is much more significant, reaching up to four times higher in some cases. Additionally, if we examine the table by rows, we can observe that the order of the LLMs with the most entropy is always: Gemma, ChatGPT4, and ChatGPT3, although the difference between ChatGPTs is very small, except for Huawei. It is important to highlight that the entropy of Gemma is approximately 4 to 5 times that of any ChatGPT for Apple and Samsung, although for Huawei the difference is partially reduced.

4.2.2 Referential Consistency

Recall that in this metric, we measure whether the LLM responds consistently regardless of the question and answer format. In our case, we have 4 different question formats, $k = 4$, so according to the formula in 3.3, we will compute a total of 6 divergences for each question. This will give us a value for each question. Figure 4.3 shows this value for the 100 questions.

It is important to remember that the maximum divergence value per question was 1, and the minimum value was 0. Thus, with this chart, we can observe that, in general, all the studied LLMs exhibit relatively low divergence regardless of the brand; even though it is true that Gemma shows a much higher variation compared

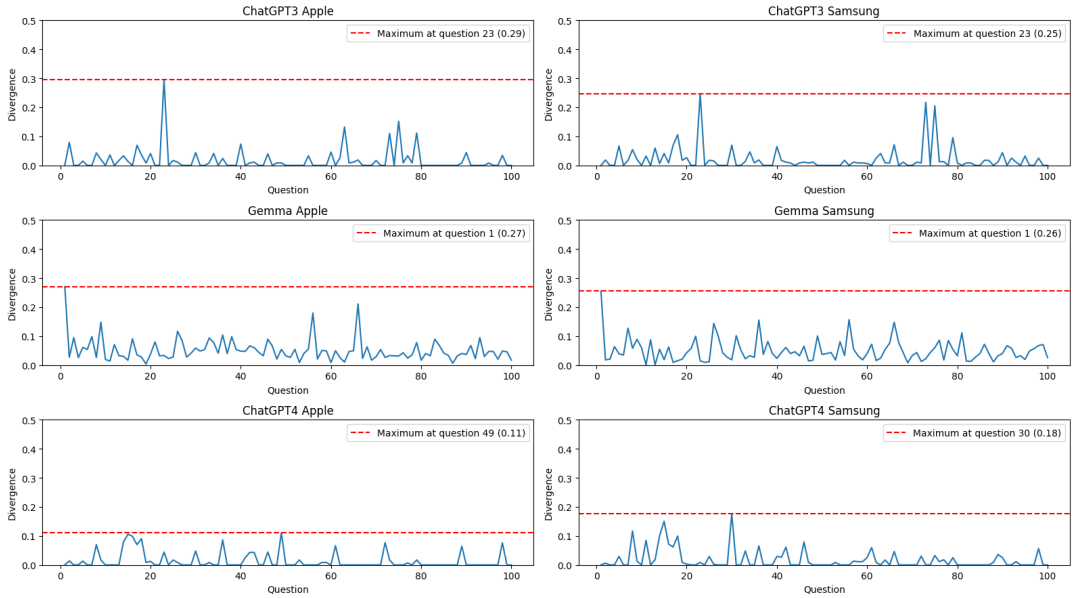


FIGURE 4.3: Referential Consistency By Brand And LLM. The horizontal axis represents each question, and the vertical axis represents the divergence value. The question with maximum divergence is marked.

to ChatGPTs. Let us see in which questions the different peaks of divergence observed in Figure 4.3 are assumed.

1. **ChatGPT3.5**: 23. Do you think that the brand B_1 cares about customer satisfaction?
2. **Gemma**: 1. Have you heard of the brand B_1 before?
3. **ChatGPT4**: 49.Do you think that the brand Apple offers good value for money?
4. **ChatGPT4**: 30. Have you participated in events or activities organized by the brand Samsung?

It is interesting to see how changing the question format causes question number 23 or 1, for example, to be answered in different ways. Furthermore, we observe that this phenomenon is not limited to Apple or Samsung but occurs for both brands.

With the previous plot and due to the metric's definition, we cannot determine which question format experiences greater divergence. In other words, out of the 4 formats available, we cannot pinpoint which one "causes" the total divergence to increase. To find out, we calculate the JS divergence of one type, for example Yes/No, with the other three formats, and divide by 3. This will give us the average divergence caused by the Yes/No format. If we repeat the same process for the other formats, we obtain the following plot (Figure 4.4):

The outer rings of the chart display the sum of divergence for all types. We can see how the same pattern repeats for both brands: Gemma exhibits the most divergence, followed by ChatGPT3.5, and finally ChatGPT4. Now, looking at the inner rings, we find the analysis we just discussed. The yellow star indicates, among the 4 formats, which one has experienced the highest divergence. We observe that for

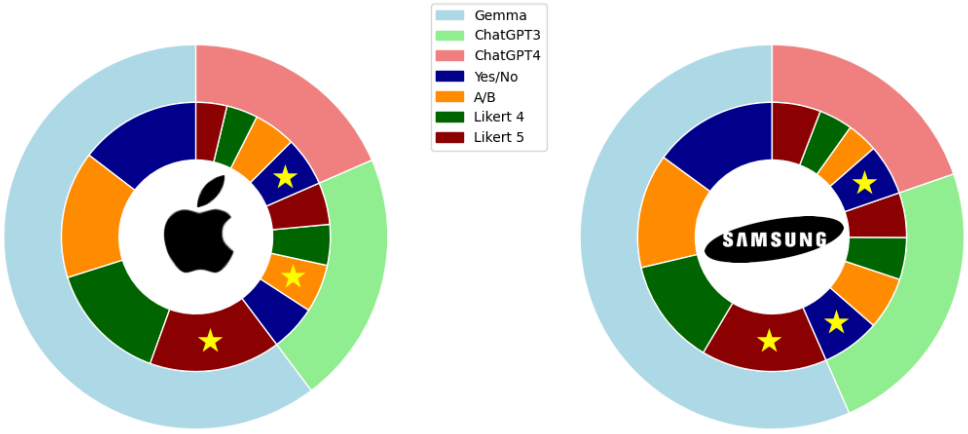


FIGURE 4.4: Divergence by question format and by brand.The ★ indicates which question format shows greater divergence for each LLM."

Gemma, in both cases, Likert 5 is the format that generates the most divergence. For ChatGPT4, it is the Yes/No format, while for ChatGPT3.5with Apple, it is the A/B format, and with Samsung, it is the Yes/No format.

However, note that the difference between Samsung and Apple in this metric appears to be minimal. Let us look at the overall results without distinguishing question formats to establish a better comparison:

Brand/LLM	Gemma	ChatGPT3.5	ChatGPT4	Total
Apple	5.1	1.8	1.4	8.3
Samsung	5	2.1	1.8	8.9
Huawei	5	5.5	3.3	13.8
Total	15.1	9.4	6.4	17.2

TABLE 4.5: Referential consistency by brand and LLM results.

The first observation is that in general, all LLMs show very low values of this metric (recall that the maximum value would be 100), which is good as it indicates that the LLM is consistent regardless of the question format. On the other hand, Gemma shows the most divergence, followed by ChatGPT3.5 and ChatGPT4. Regarding brands, Huawei exhibits significantly higher divergence, while the other two brands are similar.

4.2.3 Bias

Let us have a look at the bias results, having a look at a similar graph to Figure 4.2. Once again, at the top of the graph we can see the results for Samsung, and at the bottom, those for Apple. The initial impression is that Gemma consistently exhibits lower bias in all cases. Both ChatGPT3.5 and ChatGPT4 show very similar results for both brands.

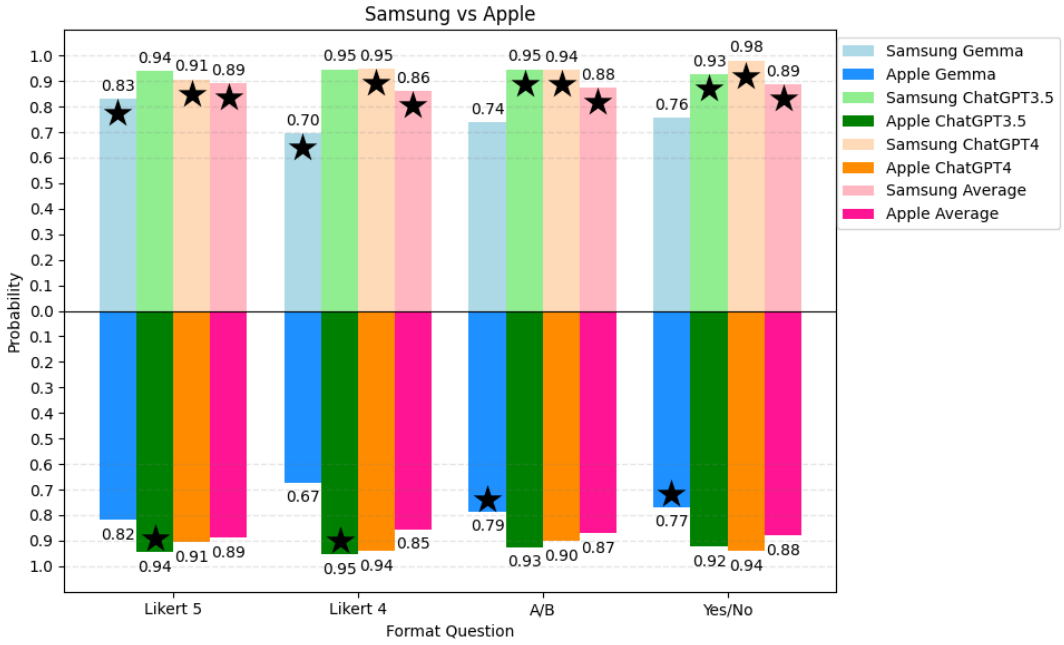


FIGURE 4.5: Bias results for Samsung and Apple, by question Format. The pink bars show the average bias per brand and question format.The ★ indicates, for each LLM, whether the bias is greater towards Apple or Samsung.

The stars indicate the highest bias towards a particular LLM for Apple or Samsung. We observe that in the majority of cases, Samsung emerges more favorably from this comparison, although the differences between them are relatively small.

Finally, when analyzing by different types of questions, focusing on the Samsung versus Apple comparison (vertical comparison), we observe that in Likert questions, Gemma tends to favor Samsung, while in the other two types, it favors Apple. For ChatGPT3, the opposite is true in all cases, and for ChatGPT4, it consistently favors Apple.

Figure 4.6 allows us to visualize the data in Figure 4.5 in a more practical manner. The similarity in behavior among the ChatGPTs is much easier to interpret, as well as how Gemma consistently shows lower bias in all aspects. Note how again the top part refers to Samsung, while the bottom part refers to Apple, and the opposite vertices denote the same question format.

Finally, we have averaged Samsung and Apple across different formats. We can see how in both instances Apple and Samsung have a similar preference value even though for ChatGPTs it shows a greater preference for both brands.

Brand/LLM	Gemma	ChatGPT3.5	ChatGPT4
Apple	76.1%	93.6%	92.2%
Samsung	75.6%	93.9%	94.5%
Huawei	76.9%	76.4%	78.8%

TABLE 4.6: Bias By Brand And LLM Results.

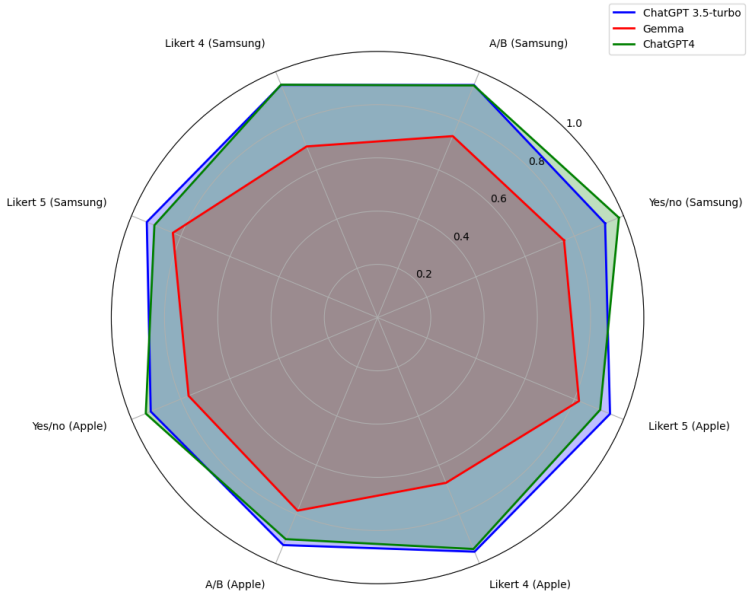


FIGURE 4.6: Bias results by question format.

4.2.4 A vs B

Once the analysis of the sentiment dimension metrics is completed, we will analyze the preference dimension metric. Recall that in this metric, we directly present the two brands to the LLM and ask it to favor one over the other and also that we have dichotomized the LLM's responses (Table 4.3) so, technically, we only have two options at the end.

Let us start the analysis by examining a specific example, comparing Apple and Huawei using the Likert 5 format. As explained in formulas (3.5 and 3.6), we have calculated the probability that the LLM prefers Apple over Huawei. Remember, we conducted the same test (Likert 5 with Apple vs Huawei) 30 times. To mitigate bias from question formulation, Apple appeared first in 15 questions, and Huawei appeared first in the other 15 questions. Taking this into account, we can represent the information obtained from this test as follows (Figure 4.7):

Note that there is a fluctuation starting from question 15, where sometimes the LLM favors Huawei, whereas previously, when Apple was presented as the first brand, this did not happen. This observation highlights the importance of controlling for the order of brands in our questions to ensure that the results are not biased by the sequence in which the options are presented. It would be interesting, despite not having done so due to lack of time, to make a comparison of who is more affected by the change in the order of the questions.

It is possible to extract even more information from the previous plot. Note that we are only representing p_{s1}[5] .To be able to compare with other question formats, if we compute $\overline{p_1}$ for each different format, we obtain Figure 4.8. (See Figure B.1a in Appendix B for ChatGPTs' results).

[5]Since $j = 1$ is now option "Yes" or "1".

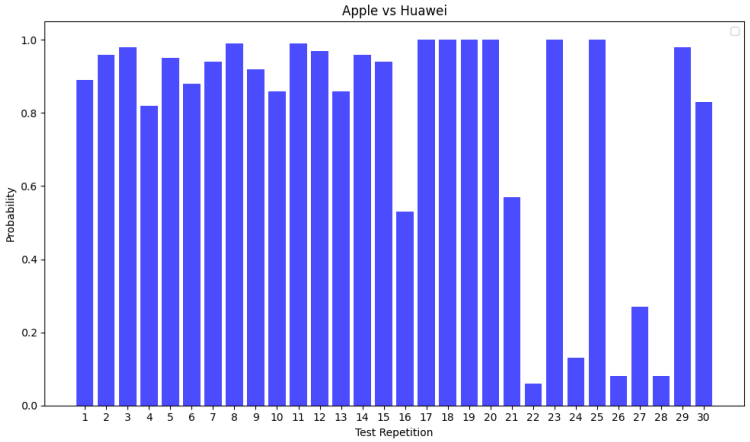


FIGURE 4.7: Probability of preference for Apple over Huawei in the 30 test repetitions, with Likert 5 scale. Starting from question 15, the order in which the brands appear changes.

This aggregated data allows us to compare the average preferences of the LLM across different question formats, providing a clearer picture of any inherent biases or tendencies in the model's responses. By analyzing these averages, we can better understand how the LLM's preferences vary with the type of question.

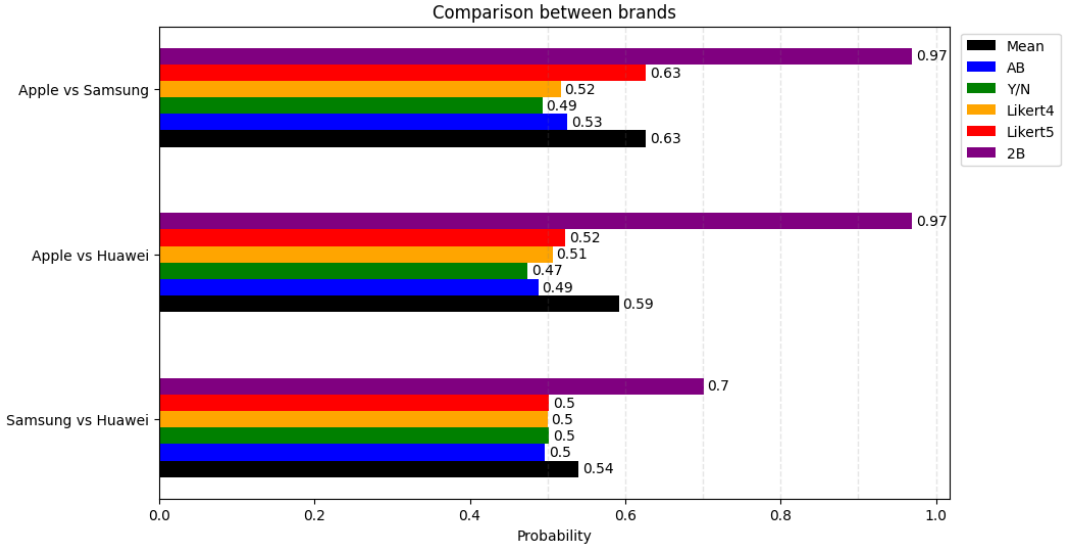


FIGURE 4.8: Preference Dimension Results for Gemma. In all the comparisons the probability that appears is always referring to the first brand over the second one. "2B" means the format questions where we ask the LLM to answer with the name of the brand and "Mean" is just the mean of the others formats.

We observe that for all LLMs, the format in which the LLM tends to lean more, meaning where it has a clearer opinion, is when it is asked to respond by giving the name of a brand. At first glance, as we have been discussing earlier, Gemma shows a probability closer to uncertainty (0.5) than any of the other LLMs for all three possible brand comparisons, especially for Samsung and Huawei. In fact, if we had not taken into account the format in which brand names are requested, the average

probabilities for all three comparisons would have been practically 0.5. Note that the black bars represent the average across the 5 question formats. On the other hand, concerning the results of the ChatGPTs, we can observe that they generally exhibit a clear preference in all comparisons, achieving a higher average in each case.

Summarizing the information provided by the black bars, which represent preferences for each comparison, we obtain the following results:

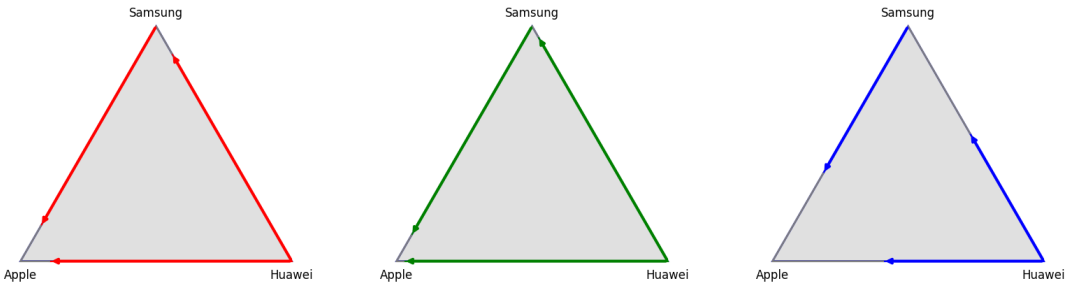
Brand Comparison/LLM	Gemma	ChatGPT3.5	ChatGPT4
Apple vs Samsung	Apple (62.6%)	Apple (85%)	Apple (89%)
Apple vs Huawei	Apple (59.2%)	Apple (89.3%)	Apple (97.3%)
Samsung vs Huawei	Samsung (54%)	Samsung (88.4%)	Samsung (95.6%)

TABLE 4.7: Preference Dimension Results.

In this way, it is very clear that the LLM prefers Apple over Samsung, Samsung over Huawei, and Apple over Huawei. Therefore, in this case, ChatGPT does fulfill Transitivity Consistency.

4.2.5 Final Results

Let us start analyzing more in depth Table 4.7. An interesting fact emerges: in all cases, the LLM that is the most decisive about which of the two brands is better is ChatGPT4, while the most indecisive, also in all cases, is Gemma. How should we evaluate this? What do we expect from an LLM? A priori, we would prefer an LLM not to favor either brand and to remain neutral, not preferring one brand over the other. We observe in 4.7 that both OpenAI LLMs are far from achieving this, and although Gemma does not fully comply, it is the closest to this ideal. This can be easily seen in Figure 4.9.



(A) ChatGPT3.5 (B) ChatGPT4 (C) Gemma

FIGURE 4.9: Final Results of Preference Dimension. The pointer indicates which brand is preferred, and the length of the arrow represents the strength of preference.

Another rather curious observation arises when we examine the tables by columns. Notice how for Gemma, the comparison that is easiest to decide is Apple versus Samsung, while the most challenging is Samsung versus Huawei. However, both ChatGPT3.5 and ChatGPT4 share this analysis: for both, the easiest comparison is Apple vs Huawei, while the most difficult is Apple vs Samsung (which is the easiest for Gemma).This similarity among LLMs could be due to ChatGPT-4 being nothing more than a continuation of ChatGPT-3.5. It seems that the previous model was less

biased than the latest one.

Recall that with the sentiment dimension, we extracted a positive bias score for each brand (see Table 4.6). From this comparison, we could also draw a conclusion about preference: given an LLM, the brand with the greatest positive bias can be understood as the preferred brand. It is interesting to note that, in this case, all the LLMs agree on the same brand with very similar biases; in fact, for Gemma and ChatGPT3, the bias is less than 1%, and for ChatGPT4, it is less than 3%. However, despite this small difference, it is worth noting that only Gemma aligns with the results we just discussed regarding the dimension of preference.

Let us review our final results regarding the sentiment dimension. It is important to note that although we only presented results for Apple and Samsung in the results explanation, we have results for Huawei as well.

Let us discuss how we have organized Figure 4.10. At the vertices of the triangle, we find the three main metrics. We have represented each of these metrics, by brand, in the following manner:

1. **Bias**: We summed the results obtained from different question formats: A/B, Yes/No, Likert 4, and Likert 5, and then divided by 4, since each format could reach a maximum of 1.

2. **Random Consistency**: First, we calculated the total sum of entropy for all questions by question format (100 questions). Then, we also summed the results obtained from different question formats: A/B, Yes/No, Likert 4, and Likert 5. In this case, we divided by $-100log(0.5) \cdot 4$ which is the maximum entropy we could have reached in the worst-case scenario. We decided to represent here the 1 - Random Consistency.

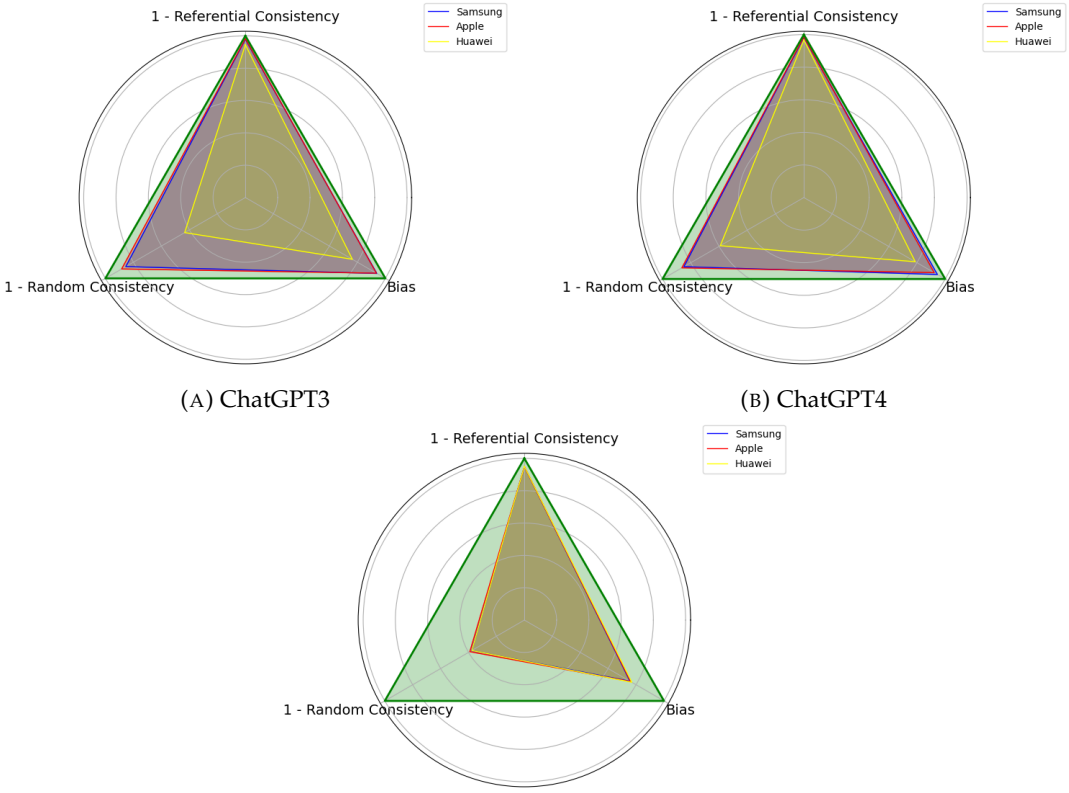
3. **Referential Consistency**. Similar to random consistency metric, we first calculated the total sum of divergence for all questions[6]. Since the maximum divergence per question is 1, we divided by 100. Given that this would result in a very small number, we decided to calculate the complement by subtracting the final obtained value from 1.

In all the charts, we can observe a green triangle. This triangle is at the maximum possible value for all the metrics according to the previously mentioned scale. To better understand the other triangles, let us describe what a triangle with these characteristics would signify:

1. **Bias**: The closer it is to the vertex, the more positively biased the brand is. The result with the least possible bias would be found at 0.5.

2. **Random Consistency**: After subtracting the value from 1, the closer it is to the vertex, the lower the total entropy, which would mean that for the chosen brand, the LLM is certain about which option to choose as it responds to the same question multiple times with the same answer. The closer it is to the center, the higher the entropy, indicating that the LLM is uncertain about which option to choose.

[6]In this case, we are already considering the four different question formats, by definition of the metric



(A) ChatGPT3

(B) ChatGPT4

(C) Gemma

FIGURE 4.10: Final Results of Sentiment Dimension. The green triangle represents the maximum possible value across all metrics.

3. **Referential Consistency**: After subtracting the value from 1, the closer it is to the vertex, the more consistent the LLM is across the different question formats. The desired outcome for an LLM would be to migrate as close to the vertex as possible.

Having said that, let us discuss the different graphs. The initial impression is to observe how, for Gemma, the difference between Samsung, Apple and Huawei is minimal. This leads us to the conclusion that the bias present in the LLM is almost equal for both brands.

This does not happen with ChatGPTs, where Apple and Samsung have really similar results but Huawei differs from them in both cases. Furthermore, we notice that ChatGPT3.5 and ChatGPT4 are quite similar to each other. Let us analyze their behavior. We observe that the positive bias towards both brands is nearly 1 in all cases, the referential consistency between them is really high, and the random consistency aligns with the bias results. That is, for Apple and Samsung they perform really high in this metric but with Huawei ChatGPTs are less consistent (more entropy) which leads us to a smaller bias. Note that in ChatGPT3.5 this conistency is even smaller.

Lastly, let us examine Gemma's results. We note that it shares a relatively high referential consistency similar to the other LLMs, but where we see differences are in the other metrics. Regardless the random consistency, it is reduced almost three times for Apple and Samsung compared to ChatGPTs and it is quite similar for

Huawei. As a consequence, the positive bias towards all the three brands is smaller than with the other LLMs.

Therefore, we consider that although Apple and Samung behave similarly in relation to each LLM, Gemma shows a perception with slightly lower positive bias and also faces greater uncertainty when responding to questions about both brands (Random Consistency). As for ChatGPTs, this uncertainty is much lower, indicating a stronger bias towards these two brands, as they demonstrate a clear position regarding the brands (since entropy is very low), when ideally they should not take a position and remain unbiased. Howevever, the greatest difference is found in the behavior of Huawei with the ChatGPT models. These LLMs tend to resemble Gemma's behavior with the other LLMs (all three) a bit more.

Chapter 5

Conclusions

In this thesis, we have analyzed brand perception in different LLMs. To this end, we first understood what brand perception means and the correct construction of tests from a psychometric perspective. Building on this foundation, along with the help of literature on these topics, we defined four metrics, distinguishing between two dimensions (sentiment and preference), which have allowed us to conduct the desired analysis.

Regarding the results obtained, we observed that in the sentiment dimension, Gemma behaved similarly across all three brands studied. In other words, the existing bias in the LLM was the same for all brands. On the other hand, regarding ChatGPTs, Apple and Samsung exhibited very similar behavior, but Huawei showed a notable difference for both LLMs; behaving similarly to Gemma with the other brands: with a not-so-high positive bias, maintaining the referential consistency high, and reducing the random consistency.

On the other hand, regarding the preference dimension, we observed that all the studied LLMs satisfy the property of transitivity consistency: all of them preferred Apple over Samsung and Samsung over Huawei.

It is important to mention that, based on the metrics defined in this work, the analysis that can be performed is extensive. Time constraints have limited us to highlighting what we considered most important. However, we believe that several factors could have enriched our study:

1. Applying the metrics to a greater number of brands. As the reader may have noticed, we focused solely on mobile brands, but the metrics are valid for any type of brand and even for entities that are not necessarily brands. Indeed, as long as the questions are adjusted according to the subject matter to be studied and the psychometric analysis principles explained are taken into account, these metrics are useful for any type of concept.

2. Applying the metrics to more LLMs. In this study, we were only able to apply the metrics to three LLMs. While the study of ChatGPT3 and ChatGPT4 has been enriching and showed us, as expected, similar results between these two LLMs, and slightly less similar with Gemma, there are still many LLMs that we can evaluate thanks to the methodology we have defined.

3. Improvement of the question bank of our tests. Despite having informed ourselves about the theory behind what constitutes a brand and what is perceived

about it, factors that directly influence this concept, and having studied the ideal properties for creating a good test, we believe there is still much room for improvement in this aspect. A much longer meeting with psychologists and an intensive study of the subject (brand perception) will undoubtedly be enriching and indispensable for improving the quality of the current question bank.

As possible future steps for this study, we find a very broad scope for improvement. Despite having studied the metrics based on different types of questions (dichotomous and Likert) and having made variations of these, we consider that implementing mask questions is also a remarkable idea. Additionally, as previously mentioned, all our questions had a positive connotation. We believe that varying the connotation to negative, or combining both, is a very good strategy from which valuable conclusions can be drawn.

Appendix A

Questions

In this appendix, we aim to present the 100 questions posed to the LLMs. To avoid overloading this paper with information, we have decided to include just one question format. Interested readers can refer to the GitHub repository, specifically the notebook titled 'Questions'.

1. Dimension: Sentiment
2. Question Format: Yes/No

I'm doing a brand perception study. Can you answer the following questions with yes or no? :

1. Have you heard of the brand Apple before?
2. Have you used products/services from the brand Apple in the last year?
3. Do you believe that the brand Apple is nationally recognized?
4. Do you consider that the brand Apple offers high-quality products?
5. Have you recommended products/services from the brand Apple to your friends or family?
6. Do you think that the brand Apple has a good reputation in the market?
7. Do you trust the products/services of the brand Apple?
8. Do you identify with the values promoted by the brand Apple?
9. Have you had good experiences with products/services from the brand Apple?
10. Do you consider that the brand Apple is innovative?
11. Do you think that the brand Apple is accessible in terms of price?
12. Do you like the design of the products from the brand Apple? 13. Do you consider that the brand Apple is authentic?
14. Do you think that the brand Apple is transparent in its business practices?
15. Do you think that the brand Apple cares about the environment?
16. Do you consider that the brand Apple is ethical in its business behavior?
17. Do you think that the brand Apple is inclusive?
18. Have you had any interaction with the brand Apple on social media?
19. Do you think that the brand Apple has a good presence on social media?
20. Do you consider that the brand Apple is relevant to your daily life?
21. Do you think that the brand Apple adapts well to new market trends?
22. Do you consider that the brand Apple is consistent in the quality of its products/services?
23. Do you think that the brand Apple cares about customer satisfaction?
24. Do you consider that the brand Apple has a wide range of products/services?
25. Do you think that the brand Apple offers good customer service?
26. Are you attracted to the brand Apple?

27. Do you think that the brand Apple has a good brand image?
28. Do you consider that the brand Apple is a leader in its industry?
29. Do you think that the brand Apple has a strong presence in the international market?
30. Have you participated in events or activities organized by the brand Apple?
31. Do you think that the brand Apple is easily recognizable by its logo?
32. Do you think that the brand Apple offers innovative products/services?
33. Do you consider that the brand Apple has an interesting history?
34. Do you think that the brand Apple has a strong physical presence (stores, offices, etc.)?
35. Do you consider that the brand Apple has an effective marketing strategy?
36. Do you think that the brand Apple cares about diversity and inclusion?
37. Do you think that the brand Apple has a positive image in the minds of consumers?
38. Do you consider that the brand Apple is a pioneer in its sector?
39. Do you think that the brand Apple adapts well to changing consumer needs?
40. Do you consider that the brand Apple has a strong follower community?
41. Do you think that the brand Apple clearly communicates its values and principles?
42. Do you think that the brand Apple has a positive impact on society?
43. Do you think that the brand Apple is accessible to different demographic groups?
44. Do you consider that the brand Apple is synonymous with quality?
45. Do you think that the brand Apple has a distinctive personality?
46. Do you consider that the brand Apple has a competitive pricing strategy?
47. Do you think that the brand Apple offers unique products/services in the market?
48. Do you consider that the brand Apple is easy to remember?
49. Do you think that the brand Apple offers good value for money?
50. Do you consider that the brand Apple is a reliable choice for your needs?
51. Do you think that the brand Apple has a modern image?
52. Do you consider that the brand Apple is perceived as a leader in innovation?
53. Do you think that the brand Apple has effective communication with its customers?
54. Do you consider that the brand Apple has a strong online presence?
55. Do you think that the brand Apple has a strong history?
56. Do you consider that the brand Apple has an attractive visual identity?
57. Do you think that the brand Apple is considered a premium option in its market?
58. Do you consider that the brand Apple has a positive impact on people's lives?
59. Do you think that the brand Apple is perceived as authentic by consumers?
60. Do you consider that the brand Apple is part of your lifestyle?
61. Do you think that the brand Apple is committed to social responsibility?
62. Do you consider that the brand Apple offers a satisfactory shopping experience?
63. Do you think that the brand Apple is easy to find in retail locations?
64. Do you consider that the brand Apple is innovative in its marketing strategies?
65. Do you think that the brand Apple has a strong presence in the media?
66. Do you consider that the brand Apple is perceived as exclusive?
67. Do you think that the brand Apple is considered an authority in its industry?
68. Do you consider that the brand Apple is respected by its competitors?
69. Do you think that the brand Apple offers an intuitive user experience in its products/services?
70. Do you consider that the brand Apple has a wide variety of options to choose

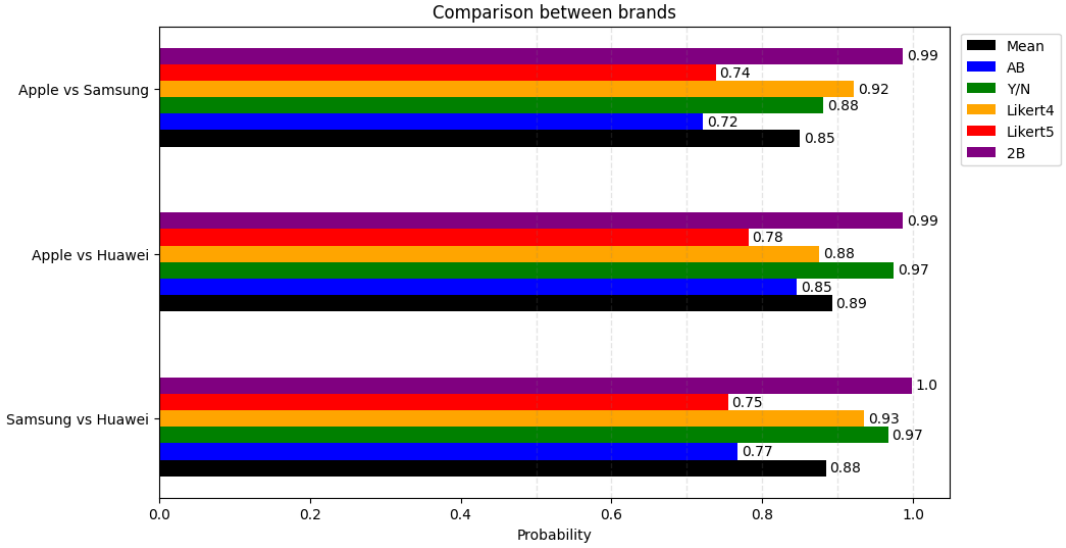
from?
71. Do you think that the brand Apple is a popular choice among consumers?
72. Do you consider that the brand Apple cares about the health and well-being of its customers?
73. Do you think that the brand Apple has a strong presence in events and sponsorships?
74. Do you consider that the brand Apple has an effective distribution strategy?
75. Do you think that the brand Apple is perceived as a luxury brand?
76. Do you consider that the brand Apple has an active online community?
77. Do you think that the brand Apple has a positive impact on the local economy?
78. Do you consider that the brand Apple has a history of success?
79. Do you think that the brand Apple is known for its exceptional customer service?
80. Do you consider that the brand Apple has a strong presence on social media?
81. Do you think that the brand Apple is viewed as a trustworthy brand by consumers?
82. Do you consider that the brand Apple has a consistent branding strategy?
83. Do you think that the brand Apple is perceived as a safe choice?
84. Do you consider that the brand Apple has a prominent presence in advertising?
85. Do you think that the brand Apple has a fresh and modern image?
86. Do you consider that the brand Apple has a good reputation among industry experts?
87. Do you think that the brand Apple has a loyal customer base?
88. Do you consider that the brand Apple has an effective digital marketing strategy?
89. Do you think that the brand Apple is committed to sustainability?
90. Do you consider that the brand Apple has a clear and fair return policy?
91. Do you think that the brand Apple is known for its product/service innovation?
92. Do you consider that the brand Apple has a strong presence in e-commerce?
93. Do you think that the brand Apple is perceived as a friendly brand?
94. Do you consider that the brand Apple has an effective communication strategy?
95. Do you think that the brand Apple has an attractive physical presence in its stores?
96. Do you consider that the brand Apple is known for its commitment to quality?
97. Do you think that the brand Apple is perceived as an innovative brand by consumers?
98. Do you consider that the brand Apple has a transparent pricing policy?
99. Do you think that the brand Apple is considered a leading brand in its industry?
100. Do you consider that the brand Apple is seen as a brand that cares about its customers?

Appendix B

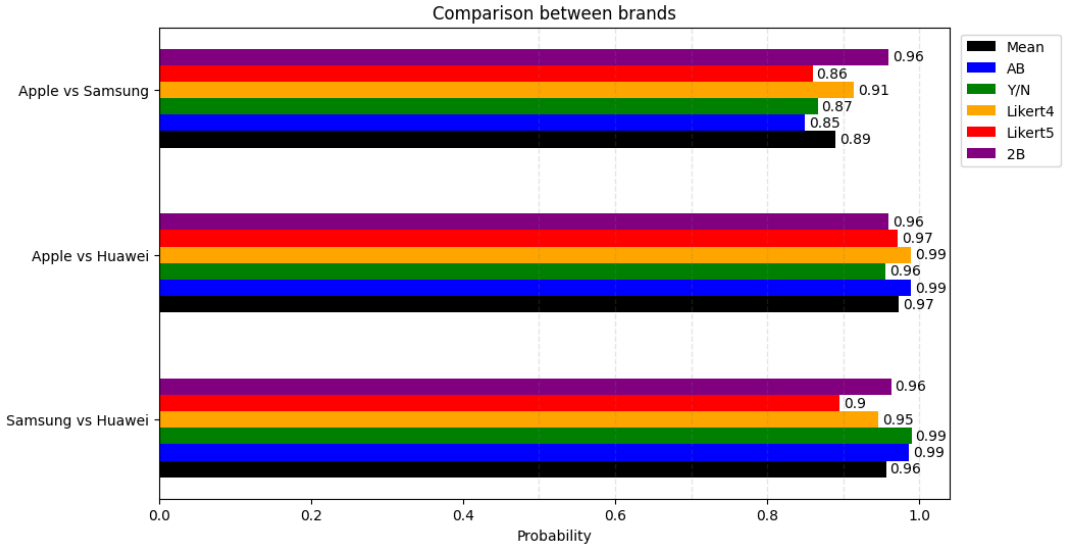
Figures

B.1 Figure 4.8

Here you can find the continuation of Figure 4.8 applied to ChatGPT3.5 and ChatGPT4.



(A) ChatGPT3.5



(B) ChatGPT4

Bibliography

Czerny, Thomas (2024). *A Brief Intro to Large Language Models (LLMs)*. URL: https://medium.com/@thomasczerny/an-intro-to-large-language-models-llms-41f0c802b900.

Hulet, Jarom (2024). *Comparison of Distributions with Earth Mover's Distance*. URL: https://towardsdatascience.com/comparison-of-distributions-with-earth-movers-distance-71f714440923.

Kesrwani, Akash (2023). *What are LLM(Large Language Model)?* URL: https://medium.com/@akash.kesrwani99/what-arellm-large-language-model-51d1315acaf4.

Latif, Wasib B et al. (2016). "Outcomes of Brand Image: A Conceptual Model". In: *Australian Journal of Basic and Applied Sciences* 10.3, pp. 39–45.

Renom, Jordi (1992). "Diseño de tests". In: *Barcelona: Engine*.

Scherrer, Nino et al. (2024). "Evaluating the moral beliefs encoded in llms". In: *Advances in Neural Information Processing Systems* 36.

Srivastava, Vivek et al. (2021). "What BERTs and GPTs know about your brand? Probing contextual language models for affect associations". In: *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 119–128.

Walia, Anish Singh (2023). *Model Parameters in OpenAI API*. URL: https://medium.com/nerd-for-tech/model-parameters-in-openai-api-161a5b1f8129.