An Analysis of Cross Cultural Typicality in Large Language Models

Nikhita Vasan, Ethan Haarer, Amogh Mellacheruvu, Sneh Gupta, May Kalnik
(nvasan7, ehaarer3, amellacheruvu3, sgupta852, mkalnik3)@gatech.edu
Georgia Institute of Technology

Abstract

Typicality effects in cognitive science describe the tendency for certain category members to be judged as more representative or "typical" of their category than others. These typicalities can give insight into how semantic knowledge is structured in the human mind, specifically with regards to lexical processing, categorization, and information retrieval. This project seeks to understand whether multilingual large language models (LLMs) align with human typicality norms. This study is a meta-analysis of typicality across 5 different languages (American English, Spanish, French, German, and Portuguese). It compares the resulting frequencies with that of a multi-lingual large language model (LLM). Prompts are based on categories that are present in all of the studies, including animals, vegetables, vehicles, fruit, clothes, sports, musical instruments, and professions. We expect LLMs' typicality judgments to align with human typicality norms in the language it is being prompted in. This study may be limited by certain inherent biases in LLMs and the obvious difficulty of representing cultural nuances solely through language. Additionally, the data collection methods, number of users, and demographics differ in each original study. Further research would focus on exploring fine-tuning LLMs on culture-specific data to increase the cultural sensitivity in typicality judgments.

A link to the code used in this paper can be found here: `https://github.gatech.edu/mkalnik3/HMLProj`

Introduction

Large language models (LLMs) represent the latest revolutionary advancement in modern technology, demonstrating human-like capabilities in many tasks including question-answering, mathematical reasoning, summarization, text generation and more. The scope of these models over recent years has grown exponentially, with the newest models boasting multilingual and multimodal capabilities designed for use in diverse and multicultural settings. Understanding how well LLMs model human cognition and represent important cultural or linguistic nuances is paramount as we sit on the precipice of an AI-driven era.

A longstanding area of research within the field of cognitive science is that of semantic categorization, or the way in which knowledge is structured and stored in the human mind. Typicality effects describe the tendency for examples of a particular category to be perceived as more reflective of their respective category than others. For example, a robin may be considered a more typical example of a bird than a penguin. The theoretical framework for this paper is based in two widely accepted theories of typicality: prototype theory and exemplar theory (Murphy, 2002).

Prototype theory suggests that any given category is represented by a "prototype" and new items are then categorized in accordance to their similarity to this prototype. For example, if robins are the prototype for the bird category, penguins would deviate from this prototype since they don't share the same characteristics (size, ability to fly, coloring etc.) and thus may not be readily categorized as a bird. While prototype theory provides a simple, efficient way to model semantic categorization, it is unable to explain contextual differences in typicality. For example, a researcher stationed in the Antarctic might consider a penguin a more typical example of a bird because that is what they encounter in their environment. Exemplar theory, on the other hand, argues that typicality is determined by the frequency of examples one has encountered in the past. New items are then categorized based on their similarity to stored exemplars. This adds an additional layer of complexity and accounts for context-dependency in typicality (Murphy, 2002).

This paper draws upon exemplar theory to test the universality of typicality judgements across diverse cultural and linguistic contexts. The goal is to determine whether a multilingual LLM (an LLM fine-tuned on large amounts of non-English data) like GPT-4o-mini can effectively replicate human typicality norms across five languages (English, Spanish, French, German, and Portuguese) and nine semantic categories (animals, vegetables, fruit, furniture, clothes, sport, musical instruments, professions). By using existing studies where native speakers produce category exemplars, this study assesses if LLMs can approximate human cognition and in what cultural or linguistic contexts they underperform.

Research Questions

1. Can multilingual large language models (LLMs) reproduce human typicality effects across different languages?

2. How effectively can LLMs reproduce typicality effects across semantic categories?

3. Can LLMs account for cross-linguistic and cross-cultural variations observed in human typicality norms?

4. Can multilingual LLMs account for language-specific inclinations in generating exemplars within typicality judgments?

Related Works

This study uses five datasets from human typicality studies conducted in the following countries/languages: Spain/Spanish, United States of America/English, Portugal/Portuguese, French/France, Germany/German. It is important to note that many of these languages are spoken globally; however this data is specifically collected from participants whose national origin is of the specified country and who are native speakers of the language. Any extrapolation

beyond these cultural and linguistic contexts cannot be justified within the scope of this study. Additionally, all of the five human studies build off of the landmark Battig study (Battig & Montague, 1969) and the subsequent updated American English study by Van Overshelde et. al. (Overshelde, Rawson, & Dunlosky, 2003). However, each paper leverages a different experimental design, which will be discussed below.

Marful et. al. expanded typicality norms for participants in Spain. 284 adult Spanish psychology students completed the standard exemplar generation task using 56 categories. Each category label was presented via computer, and participants had 60 seconds to type as many exemplars as they could (Marful, Díez, & Fernandez, 2014).

Bueno and Megherbi did the same in France, where they collected data on 70 semantic categories from over 200 adult participants. Participants were selected from two French universities and were from one of the following majors: psychology, linguistics and economics. Each participant had booklet with a page for each category and was given 30 seconds to hand-write as many exemplars as possible (Bueno & Megherbi, 2009).

Carneiro et. al. investigated category norms in Portuguese children specifically. There were three target age groups: 3-4 years, 7-8 years, 11-12 years and over 300 children participated in this study. Each group was provided a different set of categories in randomized order that were appropriate for their age. For example, the pre-school children had 13 categories while the pre-teens had 21 categories. The response time limit was also adjusted based off of age; the 3-4 years group had 90 seconds, 7-8 years had 60 seconds while the 11-12 year olds had 30 seconds to list exemplars for each category. The younger children's responses were recorded orally while the older children's responses were handwritten. This study investigated typicality from both a cross-cultural and developmental perspective, which it makes it distinct from the other papers (Carneiro, Albuquerque, & Fernandez, 2008).

Schröder et. al. conducted four different studies to expand the set of German norms. All category exemplars were first collected using the traditional exemplar generation procedure, where the researchers collected data from 20 participants (15 female, 5 male). Each participant was given a booklet with 11 category labels and asked to write as many examples as possible with no time limit. After the exemplar generation stage was complete, the remaining 140 participants participated in one of three rating studies: semantic typicality, age of acquisition and concept familiarity. The lack of a response time limit differentiates this experimental design from the others (Schröder, Gemballa, Ruppin, & Wartenburger, 2011).

Lastly, in their 2020 paper, Castro et. al. updated American typicality norms for 70 semantic categories. They use a cross-sectional sample of 246 adults in the United States split up over three age brackets: young adults (18-39 years), middle-aged adults (40-59 years), older adults (60 years+). Each participant was given a 30 second time limit to type in their responses to the categories which were presented in a randomized order (Castro, Curley, & Hertzog, 2020). Although this data includes the traditional ranking of exemplars, it also includes exemplars that were provided by participants, but were infrequent enough to not be included in the overall ranking. For the purposes of our tests, we exclude these non-ranked exemplars in favor of their ranked counterparts for accurate comparison with the other language's datasets.

We use the insights and data from these studies to develop the methodology for this project.

Methodology

Language Datasets

As discussed in the Related Works section, we leverage five existing studies on typicality in different language to assemble our datasets. Although covering many of the same topics, these datasets contain many region and culture specific categories that cannot be used, such football-related categories for the American English Dataset that didn't appear in any other study. To better mitigate such differences, we selected categories that were shared amongst all five datasets so we can standardize comparisons. This left us with nine usable categories: fruits, vegetables, clothing, sports, musical instruments, professions, vehicles, furniture and animals.

LLM as a Judge

Given the categories and exemplars provided with each language's datasets, we collected the LLM's typicality scores for each category. The model we chose gpt-4o-mini as it is fine-tuned on a multilingual corpus that includes all of the languages covered by our datasets. For each language, we prime the model to adopt the persona of a 35-year old adult of each country who is knowledgeable of the cultural and linguistic customs and experiences of a person living there. We then explain to the LLM what the typicality effect is, including examples to ensure understanding of the task at hand. We prime the model this way to ensure that every ranking produced by the model is judged consistently across all languages, and we don't rely on the predictive nature of the LLM.

This study uses the "LLM as a Judge" paradigm, in which we provide the LLM with a prompt and series of responses and ask the model to rate the responses according to the instructions. All prompts used to initialize instances of the model are first translated into their respective languages before results are collected. This translation was done independently with gpt-4o while the more computationally expensive rating portion was done using gpt-4o-mini. Prompt translation ensures that the model is still operating under the specified persona when providing its responses, which would ideally yield more congruency with the human data.

See Appendix F for additional details on prompt structure.

Evaluation Metrics

In our case, for each exemplar in a category, we ask the LLM instance to rate how typical that exemplar is of the category at

hand. We use two distinct rating scales for a more robust evaluation strategy: likert scale and traditional scale. We chose a seven-point likert scale as opposed to the five-point version to encourage more variation in model responses. The qualitative scale is as follows: Strongly Disagree, Disagree, Somewhat Disagree, Neutral, Somewhat Agree, Agree, Strongly Agree. For evaluation purposes, we convert the likert score to a numerical rating from 1 to 7.

We then do the same thing for a traditional 1 to 10 rating, with 1 representing a highly atypical exemplar and 10 representing the a highly prototypical exemplar for that category. For each exemplar in a category, we run the prompt 10 times and calculate the average likert and ranking scores for each exemplar. These exemplar-level scores are further aggregated to produce the category-wide likert and rating scores for every category in a language. We normalize the scores using Min-Max (0,1) scaling for all collected categories in each language before analyzing the data.

To compare the results collected from the LLM, we utilized the Spearman Correlation coefficient. For each language-category pair, we compute the correlation between the human responses and the model responses. When comparing across languages, we only use shared exemplars when calculating the correlation coefficient. The collected data can be found in the appendix, and the calculated spearman correlation heatmaps can be found in the following section.

Results and Discussion

Human vs GPT

Appendix A displays the results of the GPT typicality ratings and the human data. Across all languages and categories, results were mediocre. For almost all of the languages and categories, the GPT typicality rating had a better correlation with the human data than the likert scale ratings. This may be attributed to the fact that GPT had 10 options for ratings and only 7 for likert, therefore it wasn't ale to be as specific as it needed to be to correlate well with the human data. Portuguese had the worst results overall, while American English had the best. Comparing the different languages, there were some patterns among category performance among languages. For example the most of the languages were able to do decent in the fruit category and poorer in the sport category. However, some categories had no patterns across the languages. For example, American English and Portuguese had poor results for animal, while French German and Spanish had great results for the same category.

Table 1 reveals that, after averaging the the correlations among all the languages, the results remained pretty similar across all of the categories. The highest correlations was the rating correlation with fruit, and the lowest was the rating correlation of sport.

Appendix B displays the common exemplars in each category, and the Human data, GPT Likert score and rating score for each language. You can see the difference in the human data vs GPT data very clearly here. For example in the animal

Category	Likert Correlation	Rating Correlation
Animal	0.264056	0.368869
Clothing	0.266937	0.429640
Fruit	0.386939	0.509217
Instrument	0.399237	0.462054
Profession	0.250462	0.353826
Sport	0.253492	0.207905
Vegetable	0.324528	0.484667
Vehicle	0.216928	0.374248

Table 1: Average **Human vs GPT** Correlations for Categories across all Languages

category, while the human data had a lot of variance in their scores across all the exemplars and languages, the GPT models were consistently giving average or above average scores of typicality with very few exceptions. These tables make it evident that GPT consistently over estimated the typicality for most exemplars.

All of these results are indicative that chatGPT doesn't have a strong inert typicality as compared with humans. One possible cause is that a human and LLM don't retrieve memories the same way. While humans have quick 'shortcuts' to retrieve memories quickly, LLMs don't have this need.

Human vs Human

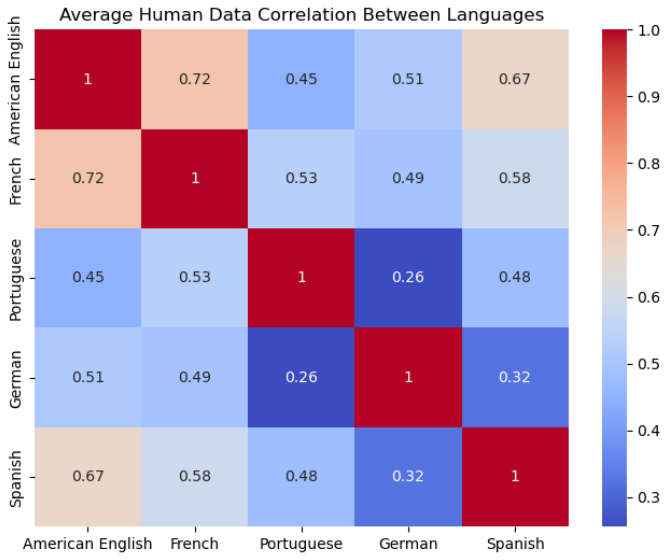

Figure 1: Average **Human vs Human** Correlations for Languages across all Categories

Next we compared how well human data correlated with human data from other languages. American English, French and Spanish had the highest correlations with each other. This implies that there is strong cross-linguistic agreement in the category structures. To be more specific as can be observed

with Figure 1, French and American English has the highest correlation (r=0.72), suggesting there is similar structure in those languages. On the contrary, German seems to have the lowest average correlation with the other languages, reflecting on its unique linguistic or cultural aspects influencing the categorization.

Appendix C has the Human Correlation Heatmaps for each category. In comparing the heatmaps of each category, the "Sport" category demonstrates the highest cross-linguistic agreement, followed by "animals". This could be generalized to say there is a universally consistent conceptualization of these categories across various cultures. Conversely, "Profession" exhibits the lowest average correlation, indicating the cultural dependence for classification in that category. However, this could be attributed to the fact that the profession category had the most exemplars across all of the languages, so even though they shared many exemplars, the frequencies would be more varied compared to the other categories.

GPT vs GPT

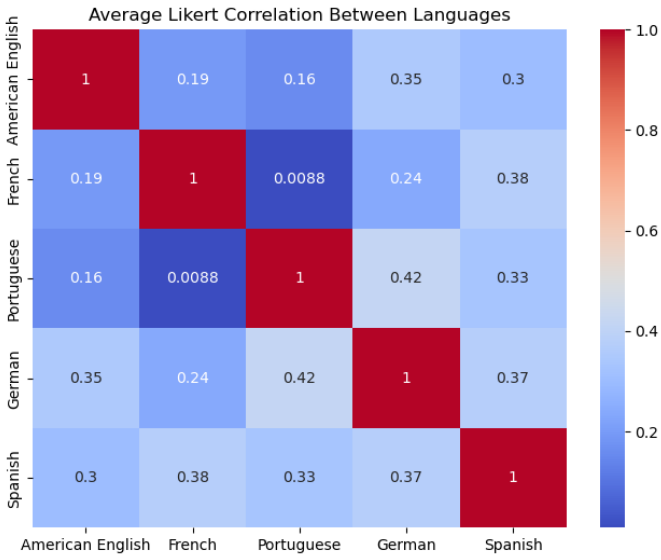

Figure 2: Average **GPT vs GPT** Likert correlations for languages over all categories

Finally we compared how well GPT data correlated with GPT data from other languages. Here, German data had the highest correlations with the other languages, most notably with Portuguese and Spanish. These results differ greatly from the human vs human data; that is, the human vs human data and GPT vs GPT data don't share the same patterns. This is an indication that not only does GPT have a weak typicality, but weak typicality regardless of culture, as it was not able to mimic the cross-cultural typicality effects in humans.

This divergence may suggest that GPT's typicality representation is mostly insensitive to linguistic and cultural context that underlies human cognition and their category norms.

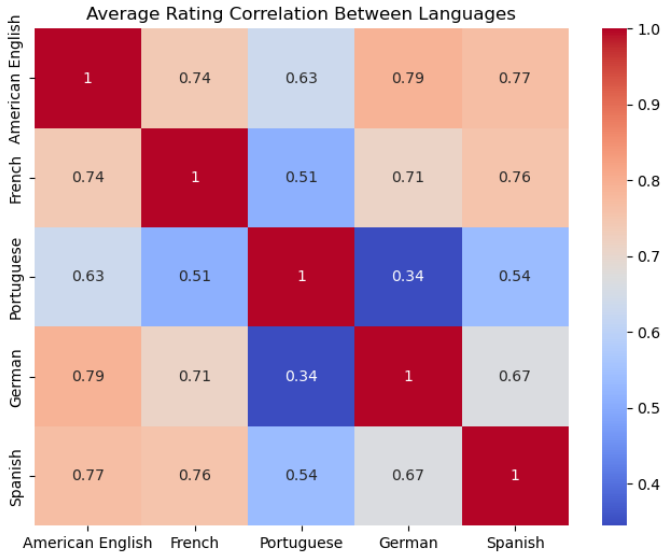

Figure 3: Average **GPT vs GPT** Numerical Rating correlations for languages over all categories

It may be worth further research into seeing how finetuning or other approaches may help to enhance the model's to better mimic human cognition. These methods are elaborated later in future work.

Appendix D and E shows the heat maps for each category when comparing the GPT vs GPT data across the languages. Vehicles performed the best, with good performance from animals, fruits and sports as well. Professions performed very poorly, likely due to the same explanation in the human vs human data analysis.

Future Directions

There are numerous future studies that can be done to better understand typicalty ratings generated by Large Language Models (LLMs) and their alignment to the human category norms.

Experimental Design

Evaluation Techniques Direct measures like generation time, which are used in human studies cannot be applied to LLMs. However, there are many methods other than the rating method used in this study to simulate human-like generation in models. Firstly, we can expand the rating and ranking approaches. We can explore alternative rating scales beyond the Likert or the 1-to-10 numerical scale. Models may have biases towards specific numbers or categories, and understanding these biases are important in analyzing the data collected. We can also avoid direct ratings altogether. Instead, we can prompt models to reorder a given set of exemplars by their typicality within a given category. We can repeat this reordering task multiple times to derive a ranking of each example.

Alternatively, models can be tasked with generating n number of exemplars in a given category. Asking the model to generate $1, 2, 3, ...n$ examples will allow us to infer implicit rankings based on generation. Lastly, we can direct the model to make pairwise comparisons between exemplars (e.g., "Is a robin or a penguin more typical bird?"). This can provide relative rankings of all examples in a category. All of these methods, direct ranking, sequential generation, pairwise comparisons, should be explored to find the best method that aligns with human generation norms. Within each method different prompts should be explored as well.

Model Priming Additionally, specializing the persona we establish for the LLM would allow for deeper and more varied analysis. Currently we have a specified age (35) and rely on the model to supply what it perceives to be the average experiences and knowledge of a given culture. Specializing a persona could help to address subtle biases introduced by general purpose multilingual training for a given model. By setting up trials to cover different age ranges, we can compare these results to existing human data, potentially granting insight into how a model changes it's perception of knowledge across different age ranges in different cultures. Specializing personas may also help to combat the dilution of cultural subtleties and nuances found in multilingual LLMs caused by sharing representations across languages.

LLM Selection In this study, a multilingual LLM was employed to analyze typicality across different languages. However, future research can also utilize monolingual models trained specifically for a single language. These models may exhibit different typicalities compared to their multilingual counterparts due to a better understanding of cultural nuance or alignment with native speakers. This will provide a better understanding of the training data's influence on category norms.

Language Choice

This study utilizes both languages spoken primarily in a single country, such as German, and languages spoken in multiple countries, such as English. Future research should investigate whether significant differences in typicality effects emerge between these language types. Additionally, there are other languages types that need to be explored including languages spoken in culturally homogeneous nations (e.g., Icelandic in Iceland) verses languages spoken in culturally diverse nations (e.g., Hindi in India). Regional variations and dialectical differences within the same language, such as Portuguese spoken in Portugal verses in Brazil, may also influence typicality ratings. Understanding these nuances is important as this reflects linguistic structure, cultural context, and regional differences in category norms.

Future studies should aim to broaden the scope of this study. We mainly focus on romance and germanic languages in this study, so expanding to other language families may also show changes in the ingrained typicalities for LLMs. We focus on five widely spoken languages, English, Spanish, French, German and Portuguese. It is essential to include more languages. Expanding to high-resource languages such as Mandarin, Hindi, and Japanese would increase the generalizability of findings across major linguistic groups. Moreover, include low resourced languages such as Tamil, Swahili and Slovenian is important to understand the capabilities of LLMs to align with underrepresented linguistic contexts. Collecting the human typicality norms for these languages may be challenging due to the limited availability of native speakers. Collaborating with native speakers and developing new datasets and benchmarks would further research in this area.

Human Datasets

Lastly, as mentioned in the Related Works section, the experimental design for the human baselines used were not standard across the languages. The studies varied in the age of participants, response time, response type (written or oral recording) and more. Future work should explore selecting studies whose experimental designs are more standardized to ensure a more robust comparison with model outputs.

Furthermore, many of the human datasets contain far more semantic categories than the ones used in this studies. Future studies can diversify the set of explored categories. We only examine common categories such as animals, professions and vehicles, but is still missing many like birds. Exploring more nuanced categories like emotions can also reveal more interesting insights into cross cultural typicality.

Conclusion

This study provides insights into both the capabilities and limitations of multilingual LLMs in modeling human typicality effects across various language and semantic categories. To recap, the research compares LLM-generated typicality judgments with human norms in American English, Spanish (Spain), French (France), German (Germany), and Portuguese (Portugal) across nine exemplar categories.

From the results, the LLM's ability to replicate the human typicality effects was varied across languages and categories with overall mediocre results. American English exhibited the strongest correlation, while Portuguese demonstrated the weakest. The 10-point rating scale outperformed the 7-point Likert scale in correlation to the human data consistently, demonstrating that a more granular scale is better aligned with human typicalities.

The human data demonstrates strong cross-linguistic agreement, specifically between American English, French and Spanish - implying similarities in categorical structure across these languages. The performance of LLM's across the categories varied, with fruit showing consistently high correlations and sports and professions demonstrating lower correlations. Based on numerical rating, the correlations between LLM-generated data across different languages were generally higher than those observed in human data, but the correlations by Likert rating were much lower. This suggests

the possibility that LLM might not have a consistent internal representation of typicality.

In terms of what it means for cognitive science and NLP, these findings emphasize the need for culturally sensitive training approaches to better understand subtleties in typicality judgments across different linguistic and cultural contexts. To briefly mention future directions, expanding the research to include a wider range of languages, including those spoken in culturally homogenous and diverse nations, as well as regional variations, offer deeper insights into the ability of the LLM. Additionally, exploring into alternative methods of eliciting typicality judgments from LLMs, like sequential generation and pairwise comparisons could offer more accurate representations of LLM-based typicality effects.

While multilingual LLMs exhibit promise in human typicality judgment replication, there is still significant room for improvement. This study lays the foundation for future research surrounding improving linguistic and cultural sensitivity in LLMs. This ultimately contributes to the development of more sophisticated language systems that can model human cognition for future research.

References

Banks, B., & Connell, L. (2023). Category production norms for 117 concrete and abstract categories. *Behavior Research Methods*, *55*, 1292–1313. doi: 10.3758/s13428-021-01787-z

Battig, W., & Montague, W. (1969). Category norms for verbal items in 56 categories: A replication and extension of the connecticut category norms1. *Journal of Experimental Psychology Monograph*, *80*(3), 1-46.

Bueno, S., & Megherbi, H. (2009). French categorization norms for 70 semantic categories and comparison with van overschelde et al.'s (2004) english norms. *Behavior Research Methods*, *41*(4), 1018–1028. doi: 10.3758/brm.41.4.1018

Carneiro, P., Albuquerque, P., & Fernandez, A. (2008). Portuguese category norms for children. *Behavior Research Methods*, *40*(1), 177–182. doi: 10.3758/brm.40.1.177

Castro, N., Curley, T., & Hertzog, C. (2020). Category norms with a cross-sectional sample of adults in the united states: Consideration of cohort, age, and historical effects on semantic categories. *Behavior Research Methods*. doi: 10.3758/s13428-020-01454-9

Marful, A., Díez, E., & Fernandez, A. (2014). Normative data for the 56 categories of battig and montague (1969) in spanish. *Behavior Research Methods*, *47*(3), 902–910. doi: 10.3758/s13428-014-0513-8

Murphy, G. L. (2002). *The big book of concepts*. MIT Press. doi: 10.7551/mitpress/1602.001.0001

Overshelde, J. V., Rawson, K., & Dunlosky, J. (2003). Category norms: An updated and expanded version of the battig and montague (1969) normsq. *Journal of Memory and Language*, *50*, 289–335. doi: 10.1016/j.jml.2003.10.003

Schröder, A., Gemballa, T., Ruppin, S., & Wartenburger, I. (2011). German norms for semantic typicality, age of acquisition, and concept familiarity. *Behavior Research Methods*, *44*(2), 380–394. doi: 10.3758/s13428-011-0164-y

Shah, R. S., Bhardwaj, K., & Varma, S. (2024). Development of cognitive intelligence in pre-trained language models. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 9632–9657). doi: 10.18653/v1/2024.emnlp-main.539

1 Appendix

1.1 Appendix A: Correlation for Each Language and Category

Category	Likert Correlation	Rating Correlation
Animal	-0.024204	0.663460
Vegetable	0.310393	0.630503
Vehicle	-0.034595	0.597832
Fruit	0.536845	0.592198
Clothing	0.018920	0.545553
Sport	0.695617	0.463671
Instrument	0.642537	0.800039
Profession	0.299980	0.572249

Table 1: Correlations for American English

Category	Likert Correlation	Rating Correlation
Animal	0.419458	0.547215
Vegetable	0.387001	0.553788
Vehicle	0.034217	0.411785
Fruit	0.272842	0.818175
Clothing	0.361280	0.502745
Sport	0.107946	0.462617
Instrument	0.475869	0.456536
Profession	0.338937	0.343899

Table 2: Correlations for French

Category	Likert Correlation	Rating Correlation
Animal	-0.141950	-0.135917
Vegetable	0.030014	0.524571
Vehicle	0.298193	0.321077
Fruit	0.178435	0.181740
Clothing	0.200858	0.537272
Sport	0.272296	-0.041367
Instrument	0.190141	0.225352
Profession	0.397794	0.441683

Table 3: Correlations for Portuguese

Category	Likert Correlation	Rating Correlation
Animal	0.633059	0.398300
Vegetable	0.511828	0.480458
Vehicle	0.723751	0.414838
Fruit	0.533625	0.550347
Clothing	0.494335	0.412346
Sport	-0.115404	-0.064372
Instrument	0.441003	0.565643
Profession	-0.011238	0.082080

Table 4: Correlations for German

Category	Likert Correlation	Rating Correlation
Animal	0.465952	0.462509
Vegetable	0.444265	0.456013
Vehicle	0.080583	0.221235
Fruit	0.519227	0.665901
Clothing	0.292424	0.299360
Sport	0.335248	0.234352
Instrument	0.364155	0.452027
Profession	0.254055	0.409140

Table 5: Correlations for Spanish

1.2 Appendix B: Shared Exemplar Data Tables

Exemplar	American English	French	Portuguese	German	Spanish
lion	0.91	0.98	0.68	0.80	0.99
elephant	0.46	0.37	0.20	0.60	0.59
cow	0.95	1.00	0.12	0.95	1.00
horse	0.31	0.27	0.24	0.55	0.61
goat	0.21	0.22	0.28	0.60	0.40
rhinoceros	0.17	0.15	0.06	0.60	0.21
dog	0.05	0.08	0.13	0.40	0.23
giraffe	0.54	0.51	0.77	0.75	0.73
pig	0.49	0.34	0.12	0.70	0.70
zebra	0.19	0.24	0.15	0.80	0.24
sheep	0.23	0.03	0.08	0.65	0.37
mouse	0.08	0.03	0.13	0.40	0.20
tiger	0.18	0.23	0.07	0.55	0.28
hippopotamus	0.43	0.23	0.11	0.55	0.61
cat	0.15	0.09	0.08	0.50	0.30

Table 6: Human Data for Animal Exemplars

Exemplar	American English	French	Portuguese	German	Spanish
lion	6.0	6.8	6.0	7.0	6.0
elephant	6.0	5.8	6.0	7.0	6.1
cow	6.9	7.0	6.2	7.0	6.2
horse	1.9	6.4	4.4	5.7	6.3
goat	6.1	7.0	3.6	5.0	3.1
rhinoceros	6.0	5.8	6.0	5.0	6.2
dog	6.0	5.0	2.8	3.0	4.1
giraffe	6.0	7.0	6.5	7.0	6.6
pig	6.0	7.0	3.2	3.8	6.6
zebra	5.5	6.9	5.2	7.0	2.2
sheep	6.0	6.3	6.0	6.9	6.2
mouse	5.5	7.0	2.3	3.2	6.3
tiger	6.2	7.0	6.0	6.4	6.1
hippopotamus	6.0	7.0	2.0	5.8	6.6
cat	6.0	7.0	3.2	5.0	6.3

Table 7: Likert Scale Averages for Animal Exemplars

Exemplar	American English	French	Portuguese	German	Spanish
lion	9.50	9.50	8.2	9.00	9.50
elephant	9.50	8.60	8.0	9.00	8.90
cow	9.95	9.50	9.2	9.40	9.50
horse	9.50	7.70	4.3	8.00	7.95
goat	7.70	6.70	5.3	7.65	7.45
rhinoceros	7.60	6.65	7.6	0.00	7.75
dog	6.92	4.50	4.6	5.20	6.78
giraffe	9.55	8.80	8.2	9.00	9.40
pig	9.52	8.80	6.9	8.20	8.65
zebra	3.95	7.00	6.8	8.00	3.35
sheep	9.00	7.65	8.1	7.15	7.55
mouse	7.75	4.07	4.5	3.90	4.55
tiger	8.30	8.30	8.0	8.00	8.50
hippopotamus	9.50	7.70	3.0	8.00	8.50
cat	7.70	7.40	5.7	6.10	6.70

Table 8: Rating Averages for Animal Exemplars

Exemplar	American English	French	Portuguese	German	Spanish
onion	0.44	0.12	0.16	0.65	0.23
broccoli	0.62	0.74	0.67	0.95	0.07
radish	0.16	0.21	0.20	0.65	0.68
potato	0.24	0.30	0.06	1.00	0.18
carrot	0.39	0.03	0.62	0.40	0.29
tomato	0.27	0.05	0.21	0.55	0.24
spinach	0.40	0.29	0.23	0.50	0.08
cauliflower	0.11	0.10	0.11	0.50	0.05
cucumber	0.15	0.17	0.20	0.50	0.51
lettuce	0.31	0.39	0.40	0.85	0.43

Table 9: Human Data for Vegetables

Exemplar	American English	French	Portuguese	German	Spanish
onion	7.0	7.0	6.0	7.0	6.0
broccoli	7.0	7.0	6.0	7.0	6.8
radish	6.0	5.6	6.0	7.0	5.9
potato	7.0	4.7	6.0	7.0	6.1
carrot	6.1	1.8	6.0	7.0	6.3
tomato	6.1	1.6	6.1	7.0	6.2
spinach	6.1	1.2	6.3	7.0	6.1
cauliflower	6.0	3.1	5.8	5.1	5.8
cucumber	7.0	5.1	6.0	7.0	6.2
lettuce	5.6	1.9	4.4	3.7	4.9

Table 10: Likert Average Ratings for Vegetables

Exemplar	American English	French	Portuguese	German	Spanish
onion	8.45	8.35	8.0	8.00	7.80
broccoli	9.30	9.00	8.0	9.00	8.35
radish	7.50	7.60	8.0	8.00	7.50
potato	8.65	8.25	8.0	8.00	8.35
carrot	9.35	8.50	8.0	7.85	7.90
tomato	8.80	8.50	8.2	0.00	8.50
spinach	9.45	8.20	8.7	0.00	9.25
cauliflower	7.45	7.50	7.0	7.00	6.85
cucumber	8.90	8.40	8.0	8.00	8.30
lettuce	7.50	4.10	7.9	6.30	8.40

Table 11: Rating Average for Vegetables

Exemplar	American English	French	Portuguese	German	Spanish
car	0.26	0.35	0.26	0.20	0.03
train	0.31	0.76	0.79	0.75	0.59
taxi	0.88	0.90	0.85	0.70	0.96
boat	0.34	0.35	0.39	0.60	0.74
bus	0.06	0.08	0.22	0.20	0.06
motorcycle	0.26	0.87	0.47	0.25	0.52

Table 12: Human Data for Vehicle Category

Exemplar	American English	French	Portuguese	German	Spanish
car	2.3	1.0	5.6	5.0	2.9
train	6.0	6.9	6.0	7.0	6.1
taxi	6.0	6.9	6.0	7.0	6.0
boat	6.0	3.3	6.0	5.1	6.0
bus	6.0	6.4	6.0	6.1	6.1
motorcycle	2.2	1.2	6.0	7.0	1.7

Table 13: Likert Average for Vehicle Category

Exemplar	American English	French	Portuguese	German	Spanish
car	7.85	7.70	5.5	7.00	8.30
train	9.05	8.50	8.4	7.75	9.05
taxi	9.60	8.80	8.1	9.00	9.15
boat	8.50	7.10	8.0	7.00	7.90
bus	9.20	8.70	8.3	8.00	8.90
motorcycle	9.35	8.65	8.0	9.00	9.50

Table 14: Rating Average for Vehicle Category

Exemplar	**American English**	**French**	**Portuguese**	**German**	**Spanish**
shirt	0.27	0.06	0.06	0.80	0.01
dress	0.37	0.39	0.37	0.60	0.57
scarf	0.39	0.42	0.13	0.85	0.38
sweater	0.23	0.15	0.14	0.20	0.40
coat	0.86	0.62	0.61	0.90	0.82
blouse	0.25	0.01	0.67	0.85	0.11

Table 15: Clothing - Human Data

Exemplar	American English	French	Portuguese	German	Spanish
shirt	6.0	4.7	6.0	7.0	5.9
dress	6.0	6.3	6.1	6.3	6.0
scarf	6.0	6.5	6.0	6.7	6.0
sweater	6.0	6.5	6.0	6.5	6.1
coat	6.0	6.6	6.0	7.0	6.1
blouse	6.2	6.1	6.0	7.0	6.0

Table 16: Clothing - Likert Average

Exemplar	American English	French	Portuguese	German	Spanish
shirt	8.50	7.45	7.9	8.00	7.35
dress	9.40	8.55	8.1	8.00	8.10
scarf	8.50	8.70	7.4	8.00	8.30
sweater	7.70	8.40	8.0	0.00	8.30
coat	9.25	8.90	8.0	8.75	8.70
blouse	8.75	8.40	8.0	9.00	8.50

Table 17: Clothing - Rating Average

Exemplar	American English	French	Portuguese	German	Spanish
pineapple	0.94	0.82	0.07	1.00	0.92
orange	0.67	0.80	0.71	0.85	0.07
mango	0.21	0.37	0.07	0.80	0.59
pear	0.23	0.16	0.16	0.60	0.32
lemon	0.18	0.35	0.72	0.50	0.23
cherry	0.07	0.09	0.14	0.50	0.76
strawberry	0.75	0.60	0.77	0.75	0.85
plum	0.35	0.40	0.23	0.85	0.80
banana	0.50	0.67	0.67	0.95	0.87
melon	0.22	0.23	0.12	0.65	0.29
peach	0.23	0.16	0.06	0.90	0.41
apple	0.45	0.57	0.32	0.70	0.71

Table 18: Fruit - Human Data

Exemplar	American English	French	Portuguese	German	Spanish
pineapple	6.9	4.6	6.0	7.0	7.0
orange	7.0	1.0	6.0	7.0	7.0
mango	6.2	2.5	6.0	6.7	6.1
pear	7.0	6.8	6.0	6.9	6.0
lemon	7.0	4.2	6.0	5.1	6.1
cherry	6.0	5.1	6.0	6.9	5.2
strawberry	6.9	6.2	5.9	7.0	6.4
plum	6.1	6.7	6.0	6.3	6.0
banana	6.7	5.9	6.0	6.0	6.2
melon	7.0	1.5	5.6	5.0	2.6
peach	6.0	5.9	6.0	5.1	5.6
apple	7.0	2.3	6.0	6.8	6.9

Table 19: Fruit - Likert Average

Exemplar	American English	French	Portuguese	German	Spanish
pineapple	9.85	9.89	8.3	9.0	9.65
orange	9.50	9.50	7.8	9.0	9.50
mango	8.50	8.15	7.3	8.2	7.60
pear	8.72	8.10	8.3	8.0	8.00
lemon	8.45	7.60	7.3	7.1	7.50
cherry	8.65	7.60	8.0	7.0	7.30
strawberry	9.40	9.30	8.4	8.8	9.20
plum	8.70	8.45	7.7	8.0	8.60
banana	8.50	8.50	7.9	8.0	8.10
melon	7.50	7.00	6.3	6.9	6.72
peach	7.60	7.40	8.0	7.0	6.65
apple	9.35	9.37	8.0	9.0	9.50

Table 20: Fruit - Rating Average

Exemplar	American English	French	Portuguese	German	Spanish
tennis	0.72	0.55	0.79	0.80	0.90
gymnastics	0.91	0.85	0.91	0.10	0.98
basketball	0.08	0.15	0.06	0.90	0.01
football	0.51	0.09	0.14	0.05	0.17
volleyball	0.15	0.47	0.18	0.05	0.35
rugby	0.52	0.54	0.38	0.05	0.78
hockey	0.17	0.38	0.45	0.05	0.46

Table 21: Sport - Human Data

Exemplar	American English	French	Portuguese	German	Spanish
tennis	7.0	5.2	5.6	7.0	6.4
gymnastics	6.5	4.3	6.2	7.0	7.0
basketball	6.0	4.0	5.4	6.0	6.3
football	7.0	4.1	4.0	3.4	4.7
volleyball	6.0	6.4	5.2	3.0	5.7
rugby	6.8	2.0	6.0	7.0	6.3
hockey	6.2	3.2	5.8	5.2	6.1

Table 22: Sport - Likert Average

Exemplar	American English	French	Portuguese	German	Spanish
tennis	9.60	8.40	5.6	9.0	9.50
gymnastics	10.00	9.50	10.0	9.7	9.70
basketball	8.50	7.60	6.7	7.0	7.65
football	8.50	3.60	5.8	7.1	6.00
volleyball	7.50	8.80	5.8	4.1	6.10
rugby	9.22	9.15	7.2	8.0	8.50
hockey	8.40	7.20	6.4	7.0	7.50

Table 23: Sport - Rating Average

Exemplar	American English	French	Portuguese	German	Spanish
trumpet	0.22	0.31	0.39	0.70	0.05
violin	0.72	0.18	0.07	0.45	0.49
harp	0.75	0.90	0.52	0.85	0.91
saxophone	0.18	0.21	0.08	0.50	0.38
cello	0.22	0.05	0.07	0.60	0.12
oboe	0.13	0.06	0.26	0.50	0.27
drum	0.32	0.32	0.76	0.10	0.13
organ	0.44	0.32	0.44	0.50	0.56
guitar	0.15	0.05	0.26	0.40	0.43
viola	0.49	0.76	0.59	0.85	0.82

Table 24: Instrument - Human Data

Exemplar	American English	French	Portuguese	German	Spanish
trumpet	7.0	5.8	6.0	6.2	6.0
violin	6.0	6.3	6.0	7.0	5.9
harp	6.6	6.9	6.6	7.0	6.2
saxophone	6.0	6.0	2.3	5.0	3.9
cello	6.2	5.4	3.4	5.0	4.7
oboe	6.0	5.9	6.0	5.2	4.8
drum	6.4	6.6	3.9	5.5	6.6
organ	7.0	7.0	5.8	6.2	5.6
guitar	6.1	4.0	5.9	3.0	5.5
viola	7.0	6.4	6.0	7.0	5.6

Table 25: Instrument - Likert Average

Exemplar	American English	French	Portuguese	German	Spanish
trumpet	7.50	7.5	7.1	7.5	7.80
violin	9.05	7.3	7.7	7.9	8.00
harp	9.50	8.6	9.4	9.0	9.45
saxophone	7.05	4.6	4.6	6.0	7.40
cello	7.10	5.6	6.1	7.0	5.75
oboe	8.20	6.7	7.0	7.1	6.35
drum	8.30	7.9	6.5	7.0	8.10
organ	8.70	8.5	7.0	8.0	7.60
guitar	7.90	6.6	6.8	5.1	7.60
viola	9.30	8.5	8.0	9.0	9.45

Table 26: Instrument - Rating Average

Exemplar	American English	French	Portuguese	German	Spanish
lawyer	0.06	0.10	0.09	0.10	0.01
engineer	0.12	0.04	0.10	0.05	0.23
nurse	0.08	0.08	0.07	0.05	0.00
doctor	0.17	0.11	0.12	0.05	0.01
teacher	0.79	0.44	0.06	0.15	0.01
dentist	0.13	0.03	0.21	0.50	0.01
mechanic	0.65	0.20	0.06	0.05	0.44
cook	0.08	0.07	0.08	0.10	0.14
carpenter	0.39	0.08	0.08	0.30	0.17
banker	0.11	0.09	0.09	0.50	0.05
secretary	0.44	0.11	0.49	0.15	0.03

Table 27: Profession - Human Data

Exemplar	American English	French	Portuguese	German	Spanish
lawyer	6.0	6.6	5.8	7.0	6.0
engineer	6.0	6.4	6.0	6.1	6.1
nurse	6.0	6.7	6.0	6.8	6.0
doctor	6.8	6.9	6.0	6.9	6.0
teacher	6.0	6.6	6.0	7.0	6.0
dentist	6.0	5.6	6.0	6.1	6.0
mechanic	6.0	5.9	6.0	6.0	6.1
cook	6.0	6.9	6.0	6.3	6.0
carpenter	6.6	6.1	6.0	6.9	6.0
banker	6.0	5.6	6.0	5.0	6.0
secretary	6.0	7.0	6.0	6.5	6.1

Table 28: Profession - Likert Average

Exemplar	American English	French	Portuguese	German	Spanish
lawyer	8.35	8.05	7.0	8.0	7.40
engineer	8.50	8.40	7.2	8.2	8.30
nurse	8.50	8.52	8.0	8.0	8.50
doctor	9.15	8.52	7.9	8.0	7.90
teacher	9.50	8.50	7.9	8.5	9.00
dentist	8.45	7.75	7.1	9.0	7.40
mechanic	8.70	4.32	7.8	8.0	8.10
cook	8.50	7.70	7.5	8.0	8.00
carpenter	9.40	7.70	8.0	8.4	8.45
banker	7.45	7.50	7.0	8.0	7.50
secretary	9.45	8.52	8.0	8.0	8.50

Table 29: Profession - Rating Average

1.3 Appendix C: Human Correlations Heat Maps

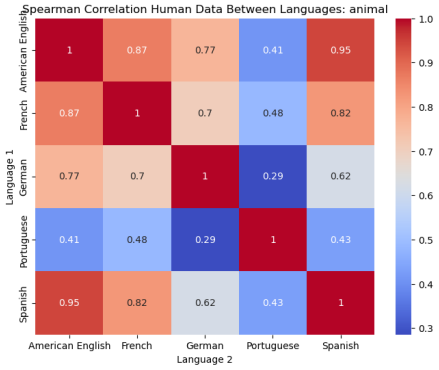

Figure 1: Heatmap showing Human correlations for the animal category.

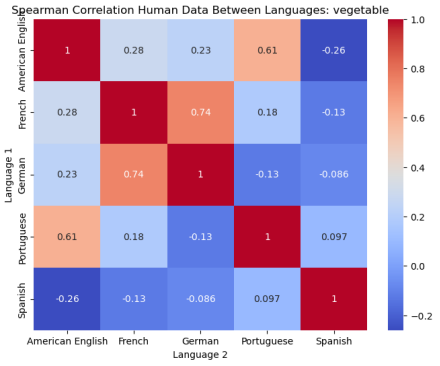

Figure 2: Heatmap showing Human correlations for the vegetable category.

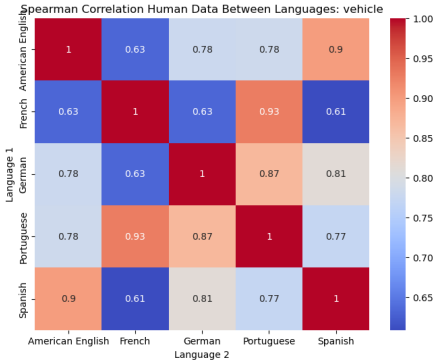

Figure 3: Heatmap showing Human correlations for the vehicle category.

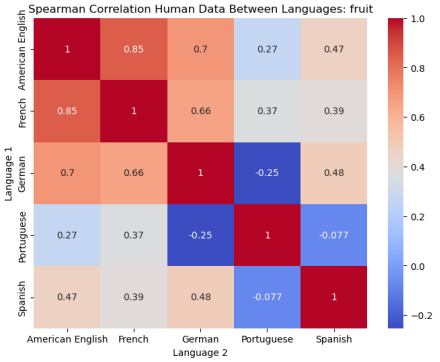

Figure 4: Heatmap showing Human correlations for the fruit category.

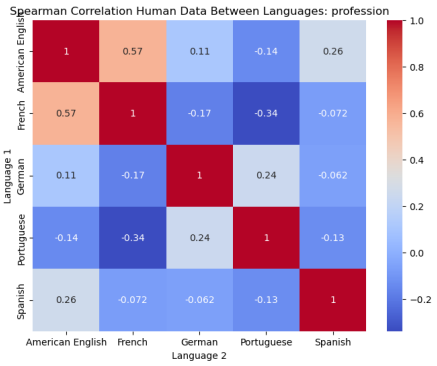

Figure 5: Heatmap showing correlations for the profession category.

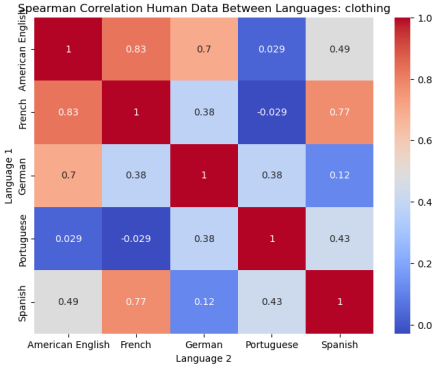

Figure 6: Heatmap showing Human correlations for the clothing category.

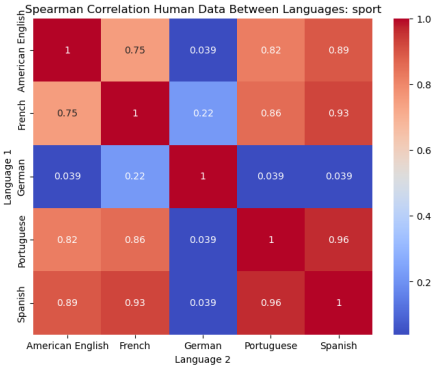

Figure 7: Heatmap showing Human correlations for the sport category.

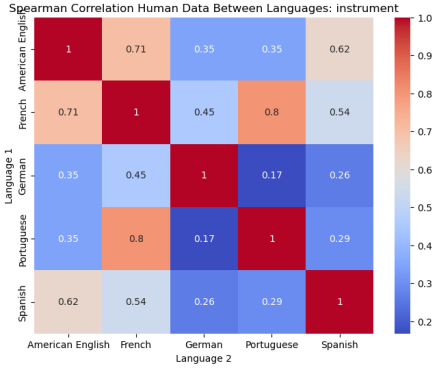

Figure 8: Heatmap showing Human correlations for the instrument category.

1.4 Appendix D: Likert Correlations Heat Maps

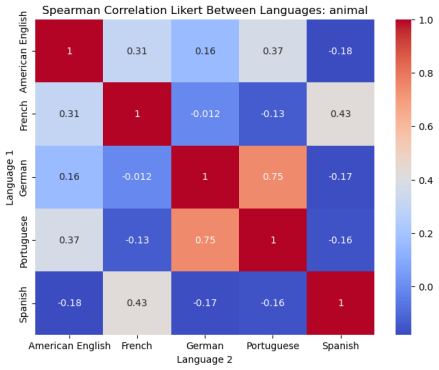

Figure 9: Heatmap showing Likert correlations for the animal category.

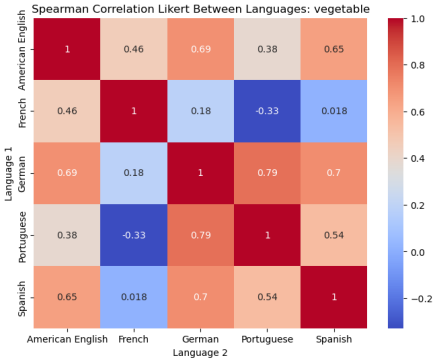

Figure 10: Heatmap showing Likert correlations for the vegetable category.

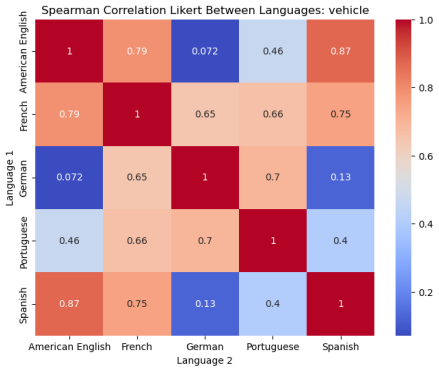

Figure 11: Heatmap showing Likert correlations for the vehicle category.

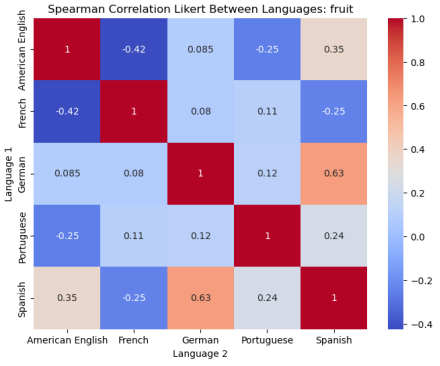

Figure 12: Heatmap showing Likert correlations for the fruit category.

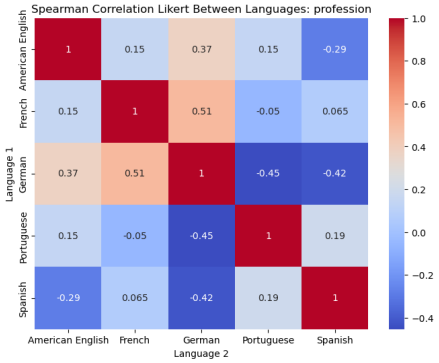

Figure 13: Heatmap showing Likert correlations for the profession category.

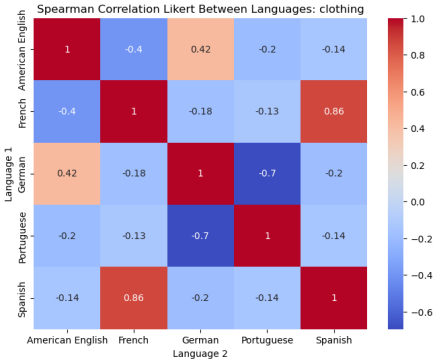

Figure 14: Heatmap showing Likert correlations for the clothing category.

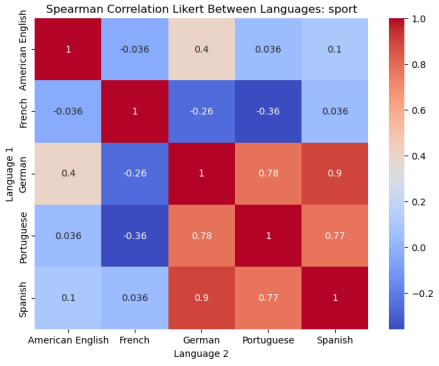

Figure 15: Heatmap showing Likert correlations for the sport category.

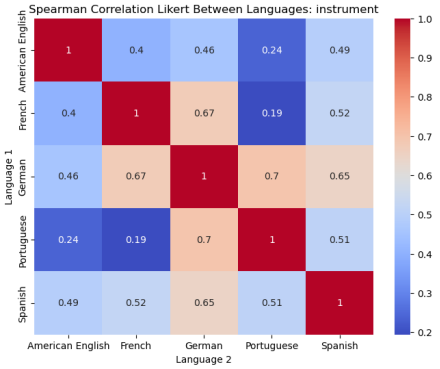

Figure 16: Heatmap showing Likert correlations for the instrument category.

1.5 Appendix E: Numerical Rating Correlations Heat Maps

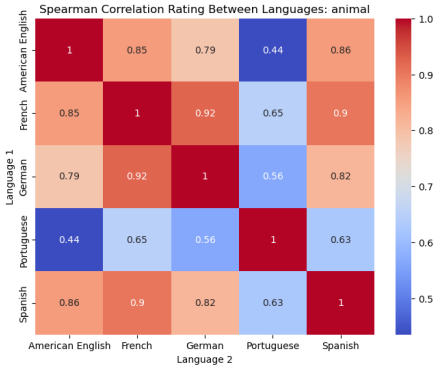

Figure 17: Heatmap of numeric rating correlations for animals.

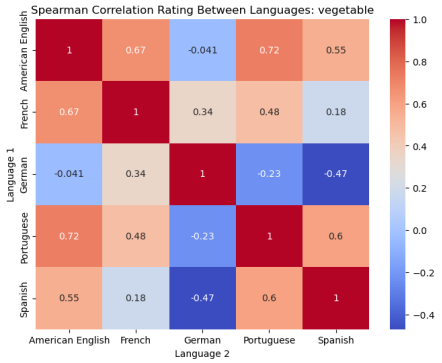

Figure 18: Heatmap of numeric rating correlations for vegetables.

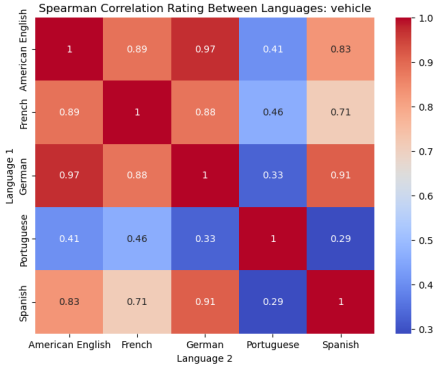

Figure 19: Heatmap of numeric rating correlations for vehicles.

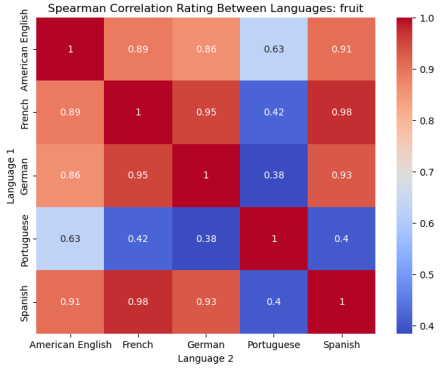

Figure 20: Heatmap of numeric rating correlations for fruit.

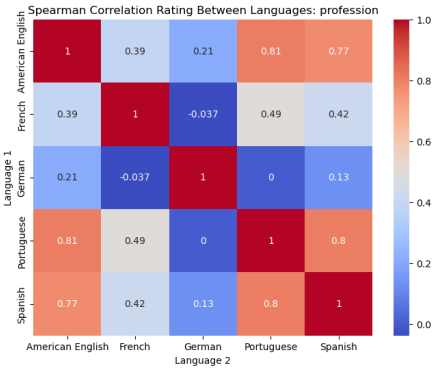

Figure 21: Heatmap of numeric rating correlations for professions.

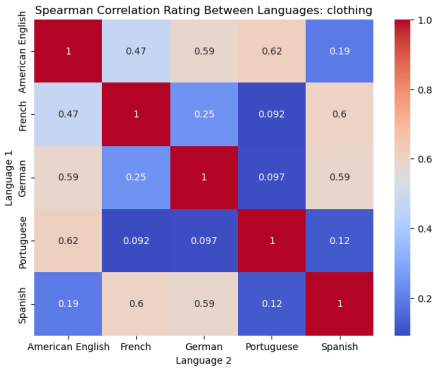

Figure 22: Heatmap of numeric rating correlations for clothing.

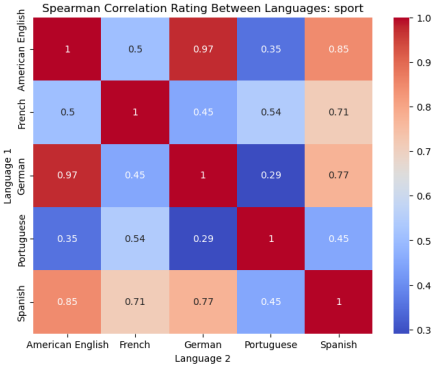

Figure 23: Heatmap of numeric rating correlations for sports.

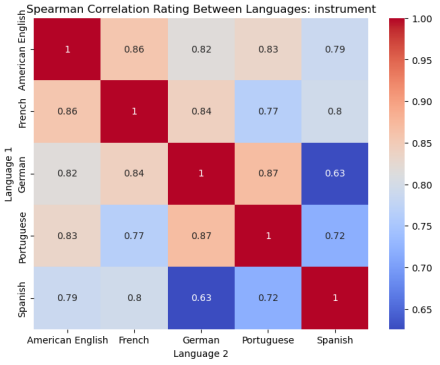

Figure 24: Heatmap of numeric rating correlations for instruments.

1.6 Appendix F: Prompt

1.6.1 Translation Prompt

System Role: You are a 35 year old adult living in {`country`}. You have a deep understanding of {`country`} culture, customs, and daily life.
User Role: Translate the following text into {`language`}, ensuring that it captures the linguistic nuances, idiomatic expressions, and cultural context of the language. Adapt phrases as necessary to make them resonate with an {`language`}-speaking audience, reflecting the tone, formality, and emotional subtleties appropriate to the situation.
`[INSERT TEXT TO BE TRANSLATED]`

1.6.2 Rating Prompt

System Role: You are a 35 year old adult living in {`country`}. You have a deep understanding of {`country`} culture, customs, and daily life. Your responses should reflect the typical experiences and perspectives of an average {`country`} person.
User Role: Typicality effects refer to the influence of the typicality or prototypicality of an object or category on various cognitive processes, including perception, categorization, and memory. The concept of typicality stems from the prototype theory, which suggests that our mental representations of categories are based on prototypes or typical examples. In the context of perception, typicality effects can influence how we perceive and recognize objects. Objects that are more prototypical or representative of a category are typically perceived more quickly and accurately than atypical objects. For example, when shown a series of pictures of birds, a typical bird like a robin would be recognized faster than a less typical bird like a penguin. In categorization tasks, typicality effects can influence how we classify objects into categories. Prototypical or highly typical objects are more likely to be assigned to their corresponding category than atypical objects. For instance, when asked to categorize fruits, an apple, being a highly typical fruit, is more likely to be classified as a fruit compared to a less typical fruit like a durian. Overall, typicality effects demonstrate how the typicality or prototypicality of objects within a category influences our perception, categorization, and memory processes, highlighting the role of prototypes in cognitive functioning.

Likert Rating: Rate how typical a {`exemplar`} is in the category of {`category`}. Use the following rating scale: strongly disagree, disagree, partially disagree, neutral, partially agree, agree, strongly agree. Only output the rating without any additional description.

Numerical Rating: Rate how typical a {`exemplar`} is in the category of {`category`}. Use a 1 to 10 rating scale, including all real numbers in the range. Only output the rating without any additional description.