Cross-Language Typicality Effects in a Multilingual Large Language Model

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

Abstract

The *typicality effect* is the finding that some members of a category are more "central" and others more "peripheral". This effect is seminal for understanding the mental representation of concepts. Recently, researchers have looked for typicality effects in the representations learned by machine learning models as evidence of their cognitive alignment. Studies of the typicality effect in Large Language Models (LLMs) have focused on models trained on English corpora and category norms collected from English speakers. Here, we use existing norms to investigate the typicality effect across five languages: English, French, Portuguese, German, and Spanish. We focused on eight categories common across these norms, and asked whether a multilingual LLM, GPT-4o-mini, shows human-like typicality effects across these languages. The results show variation in typicality gradients across languages. Importantly, GPT-4o-mini's typicality judgments show strong alignment with human norms for some languages: English and French. The strong performance for French, in particular, cannot simply be attributed to the representation of that language in the training corpus. We discuss the implications of these findings for future studies exploring alternative model prompting approaches, different languages, and the modeling of new category norms collected using uniform methods.

Introduction

Large Language Models (LLMs) have recently been shown to strongly align with human cognition across multiple domains including sentence understanding (Wilcox, 2020), mathematical thinking (R. Shah et al., 2023), and analogical reasoning (Webb et al., 2023). The scope of these models has grown rapidly in recent years, with the newest models boasting multilingual (Lai et al., 2023) and multimodal (Achiam et al., 2023) capabilities. In addition, linguistic differences are correlated with cultural and geographic differences (Huisman et al., 2019), raising the question of whether multilingual models can effectively capture variation across different cultures.

We examined how multilingual models align across languages by focusing on a foundational concept in cognitive science: *the typicality effect* (Murphy, 2002), which concerns how central or peripheral certain category members are perceived to be. This is the finding that some members of a category are more "central" and others more "peripheral" (Rosch et al., 1976). For example, a robin is a more typical member of the Bird category than a penguin. Typicality gradients have been found using two tasks. In the rating task, participants rate the typicality of exemplars on a Likert scale (Rosch et al., 1976). In the production task, participants have 30 seconds to produce exemplars of a category, and the typicality of an item is the proportion of participants who produce it (Battig & Montague, 1969).

The typicality effect can be explained by the *prototype* theory (Rosch et al., 1976).This theory proposes that each category is represented by a "prototype", which is the average of the representations of all of its members (Posner, 1968). When a new stimulus is categorized, its distance (or inversely, similarity) is computed to each prototype, and it is assigned to the concept whose prototype is closest (is most similar) (Reed, 1972). Returning to the previous example, a robin is probably more similar to the prototypical Bird than a penguin because the latter lacks many of the characteristics common to most birds: relatively small size, ability to fly (vs. swim), and so on. Thus, a robin will be rated as a more typical member of the Bird category than a penguin, and will be more likely to be produced as an exemplar of it.

Prototype theory provides a compelling account of concept representations. However, it is important to remember that typicality gradients are likely not universal. For example, a researcher stationed in the Antarctic might consider a penguin a more typical example of a Bird because they encounter penguins frequently in their environment. Prior work demonstrates that semantic alignment between languages is a product of geographical, cultural and ecological factors Thompson (2020). Therefore, if typicality gradients are the product of experience in the world, then people who live in different geographical regions may differ in what counts as typical (i.e., central) vs. atypical (i.e., peripheral) in their conceptual structures.

The goal of the current study is to investigate variation in typicality gradients across languages. This study does so in two steps. First, it evaluates whether typicality gradients do in fact vary across languages. These human data have been collected from native speakers of five languages (English, French, Portuguese, German and Spanish) who completed the production task for eight categories (Animals, Clothes, Fruits, Musical Instruments, Professions, Sports, Vegetables, and Vehicles). Second, it evaluates the cognitive alignment to the human data of a multilingual LLM, GPT-4o-mini. Specifically, this study addresses four research questions.

1. What is the variation of typicality gradients across languages (data collected in the US and Europe)?

2. Can a multilingual LLM, GPT-4o-mini, produce the same pattern of variation in typicality gradients across languages?

3. For a given language, is there alignment in the typicality gradients of human speakers and GPT-4o-mini?

4. Do languages with greater prevalence in model training data show greater human-model alignment than than languages with lesser prevalence?

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Literature Review

Language Modeling of the Typicality Effect

Initial explorations of the cognitive alignment between language models and human typicality gradients found disappointing results. Across 16 categories, the correlation between word2vec-generated typicality rankings and human typicality rankings averaged only $\rho = 0.29$ (De Deyne & Storms, 2008; Heyman & Heyman, 2019).

Recent research using modern transformer-based models has found more promising results. Misra et al. (2021) evaluated the alignment of a number of models with the ratings for 10 categories (Rosch et al., 1976), finding correlations as high as 0.40 for large variants of RoBERTa and GPT-2; see also similar work by Bhatia & Richie (2024). Most recently, Vemuri et al. (2024) evaluated the alignment between a large number of language models and the human data from 27 categories (Castro et al., 2021). Several models achieved correlations greater than 0.40, including MiniLM, MPNet, T5, and (surprisingly given its older age and architecture) GloVe.

Cross-Language Typicality Norms

Earlier, seminal studies established a standard list of categories. Battig & Montague (1969) had US participants complete the production task for 56 categories. Overshelde et al. (2003) replicated this study, again using US participants and adding an additional 14 categories, for a total of 70. The current study uses data from five recent studies that build upon Overshelde et al. (2003). They were run in different countries and recruited participants who were native speakers of different languages: United States (English), France (French), Portugal (Portuguese), Germany (German), and Spain (Spanish). It is important to note that many of these languages are spoken globally; however these data were specifically collected from participants who are native to the specified country and are native speakers of the language. Any extrapolation beyond these cultural and linguistic contexts cannot be justified within the scope of this study. We briefly describe each study and note the differences in their experimental designs.

English (US) Castro et al. (2021) collected updated US typicality norms for the 70 categories. Their cross-sectional sample consisted of 246 English-speaking adults of all ages. Each participant was given 30 seconds to type in as many exemplars of each category as possible. These data include the ranking of exemplars along with less frequent exemplars that were excluded from the rankings. In this study, we excluded non-ranked exemplars for consistency when comparing them to other languages.

French (France) Bueno & Megherbi (2009) conducted their study in France. French-speaking participants completed the production task for the 70 categories. The 200 adult participants were selected from two French universities, and were majoring in three subjects: psychology, linguistics and economics. Each participant was given a booklet with a page for each category and was given 30 seconds to hand-write as many exemplars as possible.

Portugese (Portugal) Carneiro et al. (2008) collected category norms from over 300 Portuguese children. There were three target age groups: 3-4, 7-8, and 11-12 years old. Each group was provided with a different set of categories that were appropriate for their age. For example, the preschool children had 13 categories while the preteens had 21 categories. The production time limit was also different for the groups: 90 seconds, 60 seconds and 30 seconds for the 3-4, 7-8, and 11-12 year old groups, respectively. The younger children's responses were recorded orally while the older children's responses were handwritten. The developmental nature and methodological variability of this study differentiates it from the others, a point we return to below to explain the generally weaker results for Portuguese in the current study.

German (Germany) Schröder et al. (2011) collected German norms using a production task involving 20 participants. They were each given a booklet with 11 category labels and asked to write as many examples as possible, with no time limit. The lack of a time constraint differentiates this experimental design from the others.

Spanish (Spain) Marful et al. (2015) collected category norms from 284 adults, all Spanish-speaking students majoring in psychology. They completed the production task for 56 categories. For each, the category label was presented via computer, and participants had 60 seconds to type as many exemplars as they could.

We use the category norms from these studies to address the four research questions.

Methodology

The code used in this paper can be found here: https://github.com/sneh721/Cross-Language-Typicality-Effects-in-a-Multilingual-Large-Language-Model

Language Datasets

As discussed in the Literature Review, we leveraged data from five recent studies collecting category norms in English, French, Portuguese, German, and Spanish using the production task. We selected eight categories that were shared across the datasets to standardize comparisons across languages: Animals, Clothes, Fruits, Musical Instruments, Professions, Sports, Vegetables, and Vehicles.

Procedure: LLM as a Judge

These human data consist of the typicality rankings of exemplars for eight categories across five languages. To collect these rankings from the model, we used the "LLM as a Judge" paradigm in which we provide the LLM with a prompt and ask it to generate a rating based on the instructions. Specifically, we primed the model to adopt the "persona" of a 35 year-old adult from the target country who is knowledgeable about the cultural, linguistic customs and experiences of a person living there. We also provided the definition of the

typicality effect to the model, including examples to ensure understanding of the task (R. S. Shah et al., 2024).

All prompts used to initialize instances of the model were first translated into the respective languages. This translation was done independently with GPT-4o, whereas the more computationally expensive numerical rating task was done with GPT-4o-mini. Prompt translation ensures that the model is still operating under the specified language-speaking "persona" when providing its responses, which ideally yields greater congruency with the human data. Table 1 contains a shortened version of the prompt and output.

Component	Example
Persona	You are a 35 year old adult living in {country}. You have a deep understanding of {country} culture, customs, and daily life.
Typicality Definition	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. 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.
Rating prompt	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.
Rating Output	**9.8**

Table 1: Shortened prompt and **model output** for the typicality task.

Evaluation Metrics

For each exemplar in a category, we prompt the LLM instance to rate the typicality of the exemplar in that category. We use a 1 to 10 rating scale, where 1 represents a highly atypical exemplar and 10 represents a highly prototypical exemplar for that category. The scale allows for continuous ratings, including decimal values. We run the prompt 10 times and calculate the average numerical rating for the exemplar. This is done to ensure consistency and reduce bias from individual model responses. The average ratings are then used to compute the rank order of each exemplar, with 1 corresponding to the most typical exemplar and worse ranks to less typical exemplars.

To compare the results collected from the LLM with human responses, we utilized the ranks to compute the Spearman correlations. We compute the correlation between the human typicality data and the model typicality data for each language-category pair. When comparing within a language, for each category, we use all exemplars. When comparing across languages, we only use exemplars shared across all five norms. The number of shared exemplars per category is as follows: Animals (15), Clothes (6), Fruits (12), Musical Instruments (10), Professions (11), Sports (7), Vegetables (10), and Vehicles (6). It is important to note that across languages and categories, the median number of exemplars is 49. However, when considering exemplars shared across all five norms within a category, the median drops to 10. This is a limitation of the current study.

Results

Typicality Gradients Across Languages: Human Data

Figure 1 shows the average Spearman correlation of human typicalities between languages. More precisely, each cell contains the average of eight Spearman correlations, one for each of the categories, between the typicality ratings in the two languages. By definition, then, the diagonal is one and the matrix is symmetric. Focusing on the upper triangle, the correlations between languages range from the small (0.25) to the large (0.71). The highest correlations are between English and French (0.71) and between English and Spanish (0.63). A surprising finding is the relatively low correlation between Spanish and Portuguese (0.48) given the linguistic similarity of these languages and the geographic proximity of Spain and Portugal (i.e., on the Iberian Peninsula). This may be a limitation of the Portuguese category norms, which are the only ones collected from children and not adults.

The correlations in Figure 1 are averaged across categories. A more granular look at the category level reveals interesting patterns. Summary statistics are included in Table 2.

Category	Min	Max	Median
Animals	0.37	0.94	0.71
Clothes	-0.029	0.83	0.46
Fruits	-0.2	0.85	0.42
Musical Instruments	0.1	0.82	0.41
Professions	-0.35	0.55	-0.113
Sports	-0.14	0.96	0.785
Vegetables	-0.26	0.75	0.145
Vehicles	0.49	0.89	0.77

Table 2: **Human vs Human** category level correlations.

The highest cross-language correlations were observed for Sports (0.785) followed by Vehicles (0.77). This suggests that there is general agreement across these five countries regarding these concepts. An explanation may be that globalization has resulted in shared exposure to vehicles and sports across countries (and language communities). However, this agreement may also be in part artifactual. For example, "football" refers to different sports in the US and Europe, though it is extremely popular in both regions, leading to high typicality.

Conversely, Professions (-0.113) and Vegetables (0.145) exhibited the lowest cross-language correlations, perhaps indicating the cultural specificity of these categories and corroborating Thompson's finding that semantic alignment is

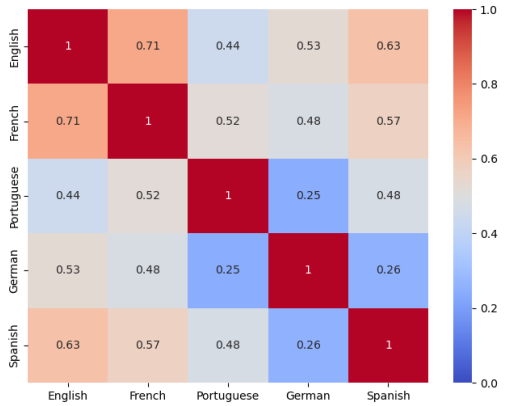

Figure 1: **Human vs Human** correlations for the five languages, averaged across the eight categories.

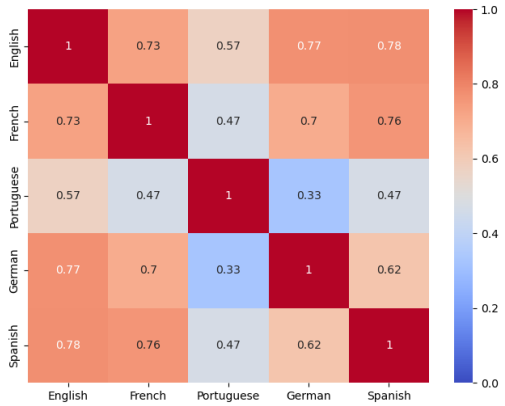

Figure 2: Average **GPT-4o-mini vs GPT-4o-mini** correlations for the five languages, averaged across the categories.

lower in culturally variable domains Thompson (2020). Alternatively, the low correlations for Professions might be an artifact of the category norms used: This category had 193 median exemplars across all five languages, but only 11 shared exemplars. For Vegetables, regional availability and dietary preferences is likely the main contributor to the variation observed.

Typicality Gradients Across Languages: GPT-4o-mini Performance

Figure 2 shows the average Spearman correlation of GPT-4o-mini's typicalities between languages. Each cell gives the the average of the eight Spearman correlations, one for each of the categories, between the model's typicality ratings in the two languages. The size of the correlations in the upper triangle range from the small (0.33) to the large (0.78). This range is comparable to that of the human data shown in Figure 1. However, the model correlations are generally higher for most language pairs with the exception of those involving Portuguese. This may reflect the lower representation of Portuguese in the GPT-4o-mini training data; see below.

We again examined the correlations at the granular level of individual categories; see Table 3.

Category	Min	Max	Median
Animals	0.51	0.91	0.79
Clothes	-0.14	0.6	0.2
Fruits	0.36	0.98	0.885
Musical Instruments	0.62	0.84	0.81
Professions	-0.009	0.75	0.38
Sports	0.36	0.96	0.5
Vegetables	-0.5	0.7	0.365
Vehicles	0.14	0.94	0.77

Table 3: **GPT-4o-mini vs GPT-4o-mini** category level correlations.

The categories that had the highest cross-language correlations included Fruits (0.885), Musical Instruments (0.81), Vehicles (0.77), and Animals (0.79). The categories that had the lowest cross-language correlations were Clothes (0.2), Vegetables (0.365) and Professions (0.38). Notably, Vegetables and Professions also had the lowest cross-language correlations in the human data. The low correlation for Vegetables might have a similar explanation as before: regional availability and the dietary preferences of speakers might greatly affect their patterns of mention in texts Thompson (2020).

The correlations in Figures 1 and 2 are analogous: the agreement of humans and GPT-4o-mini, respectively, on the typicality gradients for eight categories. We evaluated the correspondence between humans and the model by computing the Pearson correlation between the values in the upper triangle of each matrix. The correlation was of medium size ($\rho = 0.65$) and statistically significant ($p = .042$). This is evidence that GPT-4o-mini's judgments broadly capture the cross-language trends observed in the human data.

Alignment of Human and GPT-4o-mini Typicality Gradients

The primary research question concerns the alignment between human typicality ratings and GPT-4o-mini typicality ratings across languages. Figure 3 shows the average of the eight Spearman correlations, one for each category, between the human and model ratings. We focus first on the diagonal, which gives the average correlation between humans and GPT-4o-mini within the same language. These are medium/large in size for English (0.69) and French (0.74), medium in size for German (0.46) and Spanish (0.44), and small in size for Portuguese (0.17). Thus, there is alignment between humans and the model, but it varies across languages. A possible explanation for this variation is the differential representation of the languages in the GPT-4o-mini training data; we return to this possibility below.

Each off-diagonal entry gives the average correlation

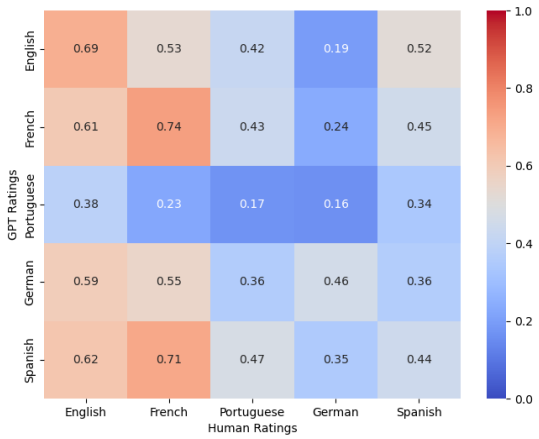

Figure 3: Average **GPT-4o-mini vs Human** correlations for the five languages, averaged across the eight categories.

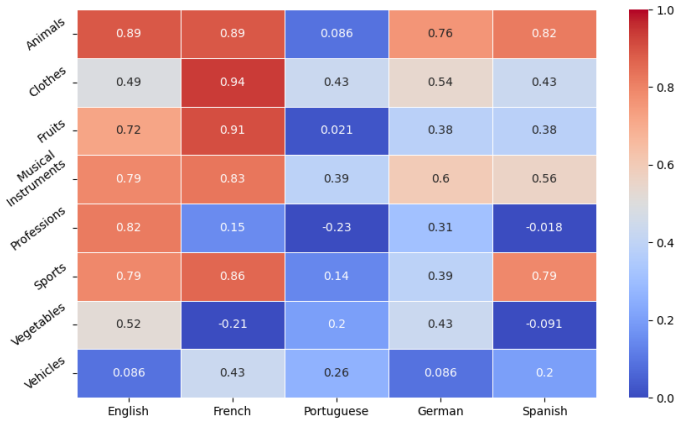

Figure 4: Average **GPT-4o-mini vs Human** Correlations per Category

across the eight categories for the typicality gradients collected from humans in one language and from GPT-4o-mini in another language. Perhaps the dominant pattern is the poor correlation between GPT-4o-mini prompted in Portuguese and human data collected in the other languages (all ρs $<$ 0.39). Additionally, there is a medium-to-high correlation between human data collected in English and GPT-4o-mini being prompted in all languages, except Portuguese (all ρs $>$ 0.58). Similarly, there is a medium-to-high correlation between human data collected in French and GPT-4o-mini output, except Portuguese (all ρs $>$ 0.52).

To better understand the moderate alignment between the human and GPT-4o-mini typicality ratings within the same language, we unpacked the averaged correlations in the diagonal entries of Figure 3. Figure 4 shows the human and GPT-4o-mini correlations within the same language separately for each of the eight categories. Across all languages, the human and model correlations are large in size for Animals, are medium/large for Musical Instruments, and are large in many cases for Sports. (The exception is again Portuguese, likely for the reasons given already.) By contrast, the human and model correlations are generally small in size (and even negative) for Professions, Vegetables, and Vehicles. (The poor alignment for the former two categories echoes our earlier findings.)

Human to GPT-4o-mini Alignment and Language Representation in the Training Data

Multilingual models are generally trained on unequal amounts of data from different languages due to resource constraints. Although OpenAI does not release the relevant information, GPT-4o is estimated to have been trained on 28% English data, with significantly smaller proportions of other languages (Hayase et al., 2024). Figure 5 plots the diagonal entries of Figure 3 – the alignment of the human and GPT-4o-mini typicality ratings for each language – against the logarithm of the estimated percentage of the training data from each language. A logarithmic scale is used to compensate for the disparity between the extreme over-representation of English in the training data compared to other languages.

Our prediction is that the greater the amounts of training data for a given language, the greater its alignment will be with human typicality ratings. This prediction received mixed support. On one hand, Portuguese, then Spanish, and finally English have increasing representation in the training data (2.3%, 2.8%, and 28%, respectively), and also show increasing average correlations between human and model typicality ratings (0.17, 0.44, and 0.69, respectively). On the other hand, German has the lowest representation in the training data (1.8%) and shows moderate alignment ($\rho = 0.46$), and French (2.9%) shows the highest alignment ($\rho = 0.74$) despite representation far below that of English.

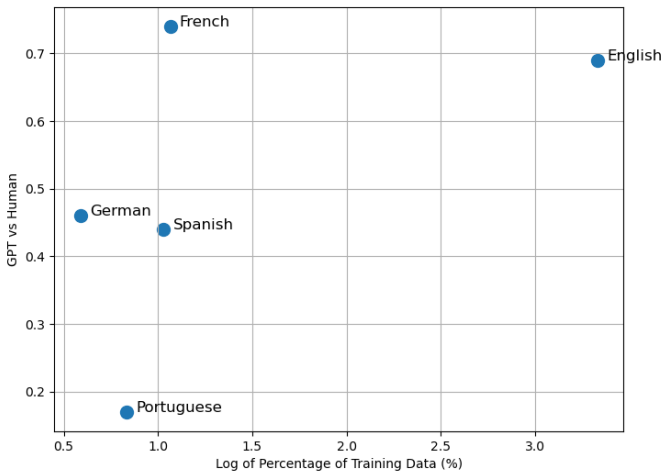

Figure 5: Diagonal of **GPT-4o-mini vs Human** correlations vs. the log of the estimated percentage of the training data.

Discussion

Exemplars of categories are organized along typicality gradients (Murphy, 2002; Rosch et al., 1976). This study has investigated the typicality effect across languages through the lens of a multilingual LLM, GPT-4o-mini. In addressing the first research question, it investigated the agreement (or lack thereof) in the typicality gradients of humans across five languages: English, French, Portuguese, German and Spanish. The correlations between pairs of languages ranged from large in size (e.g., between English and French) to small in size (e.g., between Portuguese and German). The second research question asked about the agreement in the typicality ratings of GPT-4o-mini prompted across the same five languages. The pattern of correlations was similar to that observed for humans, providing coarse evidence for the cognitive alignment of the model's concept representations. The third research question was specifically about the alignment between the human and model typicality ratings within each of the five languages. These correlations varied from large (English, French) to medium (German, Spanish) to small (Portuguese) in size. The final research question, which asked whether the size of the correlations was driven by the representation of languages in the model's training data, found mixed evidence. As a caveat, we note that GPT-4o-mini is a closed-source model. It is therefore unclear whether the collected typicalities from the model are driven solely by its exposure to multilingual datasets or whether the model has, in fact, been pre-exposed to human typicality norms in in its training dataset.

These findings show the potential and the limits of the alignment of human conceptual thinking with GTP-4o-mini's understanding of concepts. We explore these aspects and propose directions for future research.

Model Prompting Techniques In pilot work, we explored multiple prompting approaches. It is possible that additional prompting approaches might produce model typicality ratings that better align with the human data. For example, future studies might task models with generating n 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. We can also direct the model to make pairwise comparisons between exemplars (e.g., "Is a robin or a penguin a more typical Bird?") to yield relative rankings of all exemplars of a category.

LLM Selection This study employed a multilingual LLM, GPT-4o-mini, to explore typicality across different languages. A consistent finding was that cross-linguistic evaluation of LLMs must account for the fact that typicality is a culturally-laden variable. Both increasing the representation of each language in the multilingual LLM and fine-tuning the model in specific languages before collecting model ratings have the potential to improve the alignment between human and model typicality ratings. Future research can also utilize monolingual models trained on single languages. These models might exhibit different typicalities compared to multilingual models and might better capture the cultural and geographic nuances of native speakers, which is critical for both semantic (Thompson (2020) and typicality alignment.

Language Choice This study utilized both languages spoken primarily in a single country, such as German, and languages spoken in multiple countries, such as Spanish. Future research should investigate whether significant differences in typicality effects emerge between these language types. Additionally, there are other language 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 versus in Brazil, may also influence typicality ratings. Understanding these nuances is important as they reflect linguistic structure, cultural context, and regional differences in category norms. A systematic exploration would require collecting category norms across a broad range of languages and countries using a standard methodology. Such a resource might catalyze future research on cross-language typicality effects in humans and also in multilingual LLMs.

Human Datasets As noted several times above, the methodology by which the category norms were collected across languages was not standardized. The studies varied in the age of participants, times allowed for production, and response modalities (typed vs. hand-written vs. oral). On most of these dimensions, the Portuguese norms were an outlier, and this may partially explain the relatively poor alignment between the human and model typicality ratings for that language. Variation introduced by methodological differences should be eliminated in future studies by adopting a uniform procedure for data collection.

Furthermore, many of the human datasets contain far more categories than the ones used in this study. Future studies of the cross-language typicality effect should consider a larger number of categories to increase the generalizability of findings. The current study used eight categories, whereas recent norms have collected data on 70 categories in some languages (Castro et al., 2021; Overshelde et al., 2003).

Conclusion

This study provides evidence, collected across (five) languages and (eight) categories, for the potential of multilingual LLMs for modeling human typicality ratings. The strongest model-human alignment was found for English and for French, and the weakest for Portuguese. The latter result hints at limits in the available human data and representation in the model training data. Also notable was that the Professions and Vegetables categories exhibited the weakest cross-language alignment. These patterns suggest the role, beyond language, of culture and geographic region in shaping conceptual understanding.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

Bhatia, S., & Richie, R. (2024). Transformer networks of human conceptual knowledge. *Psychological review*, *131*(1), 271.

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.

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

Castro, N., Curley, T., & Hertzog, C. (2021). 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*, *53*, 898–917.

De Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 dutch words in a continuous task. *Behavior research methods*, *40*(1), 198–205.

Hayase, J., Liu, A., Choi, Y., Oh, S., & Smith, N. A. (2024). *Data mixture inference: What do bpe tokenizers reveal about their training data?* doi: 10.48550/arXiv.2407.16607

Heyman, T., & Heyman, G. (2019). Can prediction-based distributional semantic models predict typicality? *Quarterly Journal of Experimental Psychology*, *72*(8), 2084–2109.

Huisman, J. L., Majid, A., & Van Hout, R. (2019). The geographical configuration of a language area influences linguistic diversity. *PLoS One*, *14*(6), e0217363.

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Marful, A., Díez, E., & Fernandez, A. (2015). Normative data for the 56 categories of battig and montague (1969) in spanish. *Behavior Research Methods*, *47*, 902–910.

Misra, K., Ettinger, A., & Rayz, J. T. (2021). Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*.

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

Posner, K. S., M.I. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77, 353-363*.

Reed, S. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3, 382-407*.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 491–502. doi: 10.1037/0096-1523.2.4.491

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., Marupudi, V., Koenen, R., Bhardwaj, K., & Varma, S. (2023). Numeric magnitude comparison effects in large language models. In *Findings of the association for computational linguistics: Acl 2023* (pp. 6147–6161).

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

Thompson, R. S. L. G., B. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behavior, 4(10), 1029-1038*.

Vemuri, S. K., Shah, R. S., & Varma, S. (2024). How well do deep learning models capture human concepts? the case of the typicality effect. *arXiv preprint arXiv:2405.16128*.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541.

Wilcox, E. G. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.