Exploring Cultural Bias in Language Models Through Word Grouping Games

Anonymous ACL submission

Abstract

Large language models (LLMs) can exhibit cultural bias, overlooking and misrepresenting cultural nuances. Models who unequally represent global cultures can reinforce harmful stereotypes. Evaluating the extent of cultural bias in an LLM, then, is crucial to equitable model development. Most previous works focus on question-answering (QA) tasks (Palta and Rudinger, 2023). QA tasks focus on one correct answer given the cultural context, despite in many cases, there being a group of correct answers with shared characteristics for a given question. We proposed a task focusing on word groups, Word Grouping Game (WGG) that implicitly evaluates the model's cultural knowledge and norms. In WGG, LLMs are given a pool of words, where they must separate the words into groups of four words tied under a common topic. In order to perform well in the game, the model also needs to perform culture-related reasoning. We evaluated the game with two cultures, Latinx/Hispanic and Chinese, in both the native language and an English translation for comparison. Through experimentation, we find biases towards Chinese culture-based groupings, as well as disparities in performance between open- and closed-source models based on the language used for a given game.[1]

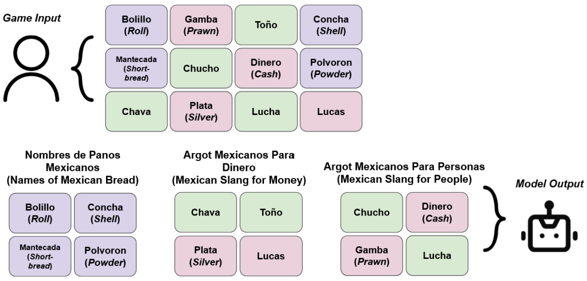

Figure 1: An example WGG game from the non-translated Latinx/Hispanic culture 3-group dataset. GPT-4 is given a pool of 12 words and is instructed to output groups of 4 words, along with the topic it believes ties each word together. In this game, GPT-4 must reason using cultural items (such as bread), as well as vernacular (such as slang and nicknames). Words of the same color belong to the same group. GPT-4 made mistakes in its grouping.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive performance in NLP tasks and are progressing toward real-world applications. Their use has expanded significantly across different continents and cultures. However, it has been shown that LLM performance drops significantly when applied across cultures, particularly from English-speaking cultures to other cultures (Shi et al., 2024; Rao et al., 2024). Beyond performance degradation, applying LLMs across cultures without considering cultural differences can lead to cultural bias, where harmful stereotypes can be formed and spread, resulting in serious consequences. Thus, identifying and measuring the degree of cultural bias in an LLM then becomes an important factor in developing equitable models.

Recent research on the evaluation of cultural bias in models has focused on explicitly soliciting cultural knowledge from the model through question-answering tasks. In these tasks, the model must identify the correct answer from a set of choices given a question in a specific cultural context (Palta and Rudinger, 2023; Rao et al., 2024; Yin et al., 2022). However, many cultural norms or biases have more than one correct answer, and these answers are often implicitly connected. For example, "bolillo" could be a valid answer to a question about typical Mexican bread, but there are also other Mexican breads, such as "concha," "manteca," and "polvoron" (as seen in Figure 1). In this paper, we explicitly test multiple connected word groups and their shared common characteristics.

To this end, we propose **WGG**, a novel word grouping game that evaluates the model's cultural

[1] We will release data and code upon paper publication.

knowledge. In WGG, the LLM is given a pool of words, and with these words, must create equal groups of four words and provide a topic that connects each group. The model also needs to reason about the given set of words to succeed because of the game's constraints, i.e., only four words per grouping. This involves optimizing the word grouping by identifying unique commonalities among the four words without interfering with other grouping choices. Figure 1 presents a Latinx/Hispanic WGG example. The dataset includes two cultures, Latinx/Hispanic and Chinese. We experimented with different difficulty level of the game. For models, we evaluate two closed-source and two open-source models. Although the open-source models were trained using different sources of data, experimentation clearly groups them together relative to the closed-source models. Open-source models performed significantly better with an English representation compared to Spanish and Chinese, while closed-source models performed more evenly across the three languages evaluated. Additionally, we find trends with subject matter, with experiments on 'Everyday' tag-specific games significantly outperforming experiments on 'Pop culture' and 'Linguistic tag-specific games.

2 Dataset Creation and Analysis

2.1 Word Grouping Game (WGG)

WGG is designed to link words under topics with colloquial and non-colloquial origins. Players must group a multiple of four words into groups of four and identify a common topic that unites them.[2] Each word can only be used once, requiring the player to consider each word's relation to others.

2.2 Word Group Collection

A dataset of Spanish word groupings centered around Latinx/Hispanic culture and a dataset of Chinese word groupings centered around Chinese culture were collected for usage in experimentation and analysis.[3]

There are a total of 48 Spanish word groupings and 80 Chinese word groupings. While these groupings do not represent the totality of each of these cultures, through focusing on aspects such as vernacular, pop culture, and local knowledge, the authors believe these groupings represent an inclusive sub-sample. These groupings were then translated literally into English. More details about translation is in Appendix B.1 This creates 4 different word group datasets for use in experimentation. The translation of the same dataset is used to analyze performance variations from different knowledge presentations.[4]

Percent Composition of Primary Tags By Culture Dataset

Primary Tag	**Culture Dataset**			
	Spanish	Example Topic (Translated)	Chinese	Example Topic (Translated)
Everyday	37.50%	Aquatic Animals	58.75%	Study Subjects
Pop Culture	31.25%	Characters from Don Quixote	21.25%	CBA League Team
Linguistic	31.25%	Words For "Cool"	20.00%	Classification of Chinese Poetry Forms

Table 1: Word group tag example and distribution. For both Spanish and Chinese word groups, We have a balanced distribution of primary tags, indicating the diversity of the collected word groups.

2.3 Word Group Tag Annotation

Following the collection of word groups, each group was tagged with a primary tag and optionally with sub-tags that would denote finer specificity in a given topic. The composition of each primary tag per culture dataset can be seen in Table 1. Different tags indicate diverse cultural knowledge in word groups. This is particularly useful if the datasets used to train the models are biased towards a subset of topics. Here are three primary tags:

Everyday denoting topics relating to common knowledge or human experience. Groupings with this topic pertain to physical objects, color, weather, emotions, etc. – all facets of life that people have come to know by living "every day".

Pop culture denoting topics relating to the many facets of modern pop culture. Groupings with this topic pertain to subjects like movies, music, etc.

Linguistic denoting topics relating to the linguistic structure or origin of words. Groupings with this topic pertain to subjects like word structure and word origin, as well as slang and synonyms.

2.4 Game Creation

The collections of word groups, i.e WGG games were used throughout experimentation. These were created by randomly sampling without replacement

[2] It is possible for a player to achieve the correct groupings, but this could be accomplished through strategies such as the process of elimination, which would not be the desired skill we aim to evaluate.

[3] Please find more details in Appendix B.1

[4] Proper nouns, such as last names, were used as is. Fill-in-the-blank topics had the remaining word segment translated if possible, and if not, kept the same.

2-Group Game Results

Models	**F1 Score**				**FastText % Topic Achieved (Threshold = 0.3)**			
Languages	S	TS	C	TC	S	TS	C	TC
GPT-3.5	0.92	0.93	**0.95**	0.88	0.71	**0.77**	**0.77**	0.73
GPT-4	0.95	0.96	**0.98**	0.96	0.77	**0.83**	0.80	0.81
LLaMA-7b	0.14	0.14	0.00	**0.22**	0.05	**0.11**	0.00	**0.11**
Mistral-7b	0.76	0.93	0.87	**0.94**	0.38	**0.80**	0.60	0.71

3-Group Game Results

Models	**F1 Score**				**FastText % Topic Achieved (Threshold = 0.3)**			
Languages	S	TS	C	TC	S	TS	C	TC
GPT-3.5	0.88	0.87	**0.93**	0.85	0.68	0.68	**0.76**	0.69
GPT-4	0.92	0.92	**0.97**	0.93	0.75	0.77	**0.78**	0.73
LLaMA-7b	**0.16**	0.15	0.00	0.05	0.05	**0.09**	0.00	0.03
Mistral-7b	0.79	0.85	0.88	**0.93**	0.35	**0.71**	0.59	0.63

4-Group Game Results

Models	**F1 Score**				**FastText % Topic Achieved (Threshold = 0.3)**			
Languages	S	TS	C	TC	S	TS	C	TC
GPT-3.5	0.82	0.83	**0.89**	0.81	0.65	0.67	**0.76**	0.66
GPT-4	0.90	0.90	**0.93**	**0.93**	0.66	0.73	**0.79**	0.68
LLaMA-7b	**0.21**	0.20	0.04	0.08	0.01	**0.10**	0.00	0.03
Mistral-7b	0.62	0.87	0.80	**0.92**	0.22	0.64	0.56	**0.57**

Table 2: Depicting the averaged results by model across each game used during experimentation. Results are divided by the number of groups in each game, and by each dataset. S denotes the Latinx/Hispanic dataset, TS the English-translated version, C the Chinese dataset, and TC the English-translated version.

2, 3, and 4 groups (tested game sizes) from each word group dataset, creating 2-, 3-, and 4-group game datasets. Please find more details about game creation in Appendix B.2

Games between four groups are naturally the most challenging, while games between two groups are the easiest. The more word groups are present in one game, there exists more possible solutions for the final solution, making the game more difficult. One could increase the difficulty of WGG by manipulating the number of groups in a game.

3 Experiments

3.1 Evaluation

Identifying Attempts Since there is no inherent order to the way the LLM answers the game, it is necessary to match the groups in model predicted groupings with the true groupings. To do this, attempted groupings are assigned to the true groupings that have the greatest set intersection with them - both groups are then pruned from their respective collections to ensure no attempt or true grouping is assigned to more than one group of the other type.

Membership Evaluation We use F1 to compare predictions against ground-truth word groups. A positive is defined as a word present in an attempted group, while a negative is a word not present in an attempted group.

Topic Evaluation Along with the predicted groupings, we also need to evaluate topic similarity. Topic similarity is evaluated through a "Topic Achieved" score (TA), a boolean denoting whether a topic was successfully guessed or not. TA is calculated by relating the FastText embeddings-based cosine similarity between a given predicted topic to its matched true topic, and the other true topics in the same game. If the predicted topic and its matched true topic have a similarity of at least 0.3 and is greater than the similarity between the predicted topic and each other true topic, a TA score of 1 is given. Details into determining using FastText and a threshold of 0.3 can be seen in Appendix C.

3.2 Baselines

We experimented with zero-shot prompting. The prompts include the rules of the game, as well as providing the pool of words for the given game. Within the prompt, the output format is specified to support the parsing of LLM answers. We experimented with different prompts for the model using each game in each respective 100-game and 50-game test dataset, and reported the results on the test subset. [5].

Models For the closed-source LLMs, we utilized GPT-3.5 Turbo (Brown et al., 2020) and GPT-4 (OpenAI et al., 2024). For the open-source LLMs, we utilized LLama2-7B-chat (Touvron et al., 2023) and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023). These two models are two of the highest performing open-source LLMs, with good performance in various NLP tasks such as reasoning, mathematics, and code generation (Jiang et al., 2023)[6].

4 Results

Performance with Culture of Origin Utilizing WGG, we are able to note significant differences in performance relative to culture. As we can see from 2 in the F1 score column, models are more biased toward Chinese culture for GPT-3.5 Turbo, GPT-4, and Mistral-7b, as evidenced by higher per-

[5] During this process, if any game caused issues with answer parsing, the game is saved and later retested to ensure an equal number of evaluations for each dataset for each model. Please find the exact prompts used for each model in Appendix D

[6] More model details are shown in Appendix E

4-Group ’Everyday’ Tagged Game Results

Models	**F1 Score**				**FastText % Topic Achieved (Threshold = 0.3)**			
Languages	S	TS	C	TC	S	TS	C	TC
GPT-3.5 Turbo	0.88	0.90	**0.94**	0.84	0.78	0.80	**0.81**	0.76
GPT-4	0.95	0.96	**0.98**	0.94	0.79	**0.86**	0.83	0.74
LLaMA-7b	**0.09**	**0.09**	0.03	0.08	0.00	**0.11**	0.00	0.04
Mistral-7b	0.29	0.44	0.88	**0.89**	0.31	**0.69**	0.66	0.61

4-Group ’Pop Culture’ Tagged Game Results

Models	**F1 Score**				**FastText % Topic Achieved (Threshold = 0.3)**			
Languages	S	TS	C	TC	S	TS	C	TC
GPT-3.5 Turbo	0.80	0.84	**0.87**	0.82	0.65	0.62	**0.81**	0.67
GPT-4	0.90	0.91	0.90	**0.97**	0.71	0.71	**0.85**	0.77
LLaMA-7b	**0.05**	0.01	0.03	0.03	0.00	**0.03**	0.00	0.01
Mistral-7b	0.19	0.33	0.71	**0.89**	0.11	0.42	0.52	**0.58**

4-Group ’Linguistic’ Tagged Game Results

Models	**F1 Score**				**FastText % Topic Achieved (Threshold = 0.3)**			
Languages	S	TS	C	TC	S	TS	C	TC
GPT-3.5 Turbo	0.70	0.69	**0.78**	0.71	0.43	0.48	**0.55**	0.39
GPT-4	0.82	0.80	**0.86**	0.82	0.40	0.51	**0.59**	0.38
LLaMA-7b	0.07	**0.08**	0.05	0.08	0.03	**0.06**	0.01	0.03
Mistral-7b	0.23	0.40	**0.75**	0.69	0.11	**0.57**	0.48	0.33

Table 3: Depicting the averaged results across each tag-specific 4-group game used during experimentation by dataset. S denotes the Latinx/Hispanic dataset, TS the English-translated version, C the Chinese dataset, and TC the English-translated version.

formance in the C and TC datasets, compared to the S and TS datasets. LLaMA-7b has an inverse relationship across these tables, having higher performance in Latinx/Hispanic culture groupings. Topic similarity clearly separates the open- and closed-source models, with the closed-source models having consistently higher Topic Achieved scores with Chinese culture, compared to the open-source models performing better with Latinx/Hispanic culture. By seeing the different effects of the culture of origin on the F1 score versus the Topic Achieved score, we see that game reasoning, achieving a correct final grouping (evaluated through the f1 score), and group reasoning, achieving a correct topic (evaluated through the topic achieved score), are deferentially impacted by the culture of origin of the word groupings. Bias towards one culture versus another can point to bias during training towards data sourced from one ethnic/cultural group over another, regardless of the language of the text itself.

Performance with Language of Origin Experimentation with WGG also exposed differences in performance relative to the language/knowledge representation. Considering the F1 score, we can see in 2 that Latinx/Hispanic culture groupings had better performance when translated to English for GPT-3.5 Turbo, GPT-4, and Mistral-7b. Chinese culture groupings had a differing relationship, with better performance in Chinese rather than English in closed-source models, but English better than Chinese in open-source models. Considering the Topic Achieved score, we can see balanced performance across the different languages for the closed-source models – as can be seen in 2, GPT-3.5 Turbo and GPT-4 score well in both S, TS, and C datasets. The open-source models, however, fared better with English knowledge representations, with better performances in TS and TC.

Performance with Game Size By varying the number of groups during experimentation with WGG, we see patterns in performance relative to game size dividing closed- and open-source models. As previously predicted, the closed-source models degrade in performance relative to the number of groups in a game – as the number of groups increases, performance in both F1 and Topic Achieved scores decreases. Open-source models, however, have inconsistent performance relative to game size, with irregular patterns in 2 in F1 and Topic Achieved score.

Performance with Tag Composition Performance differences relative to the tag-specific games show bias in subject matter. Open-source models, for both F1 and Topic Achieved score, performing far better in the ’Everyday’- and ’Pop culture’-specific games compared to ’Linguistic’-specific games, as seen in 2. Open-source models follow a different pattern, with better performance in ’Everyday’-specific games compared to all other variations, and ’Linguistic’-specific games compared to ’Pop culture’-specific games.

5 Conclusion

We have presented a new culture-based dataset, approaching QA to assess cultural bias through a novel, game-based perspective. By having game mechanics that require high performance in both reasoning and cultural knowledge, we have created an evaluation method that assesses culture-based reasoning. As evidenced through experimentation, WGG can be used to delineate bias relative to culture, language, and subject matter, noting consistent performance differences between the closed- and open-source models that were tested.

6 Limitations

The compiled dataset is not representative of the culture due to the limited number of annotators. We will expand this part of the work in the future. We also only covered two cultures, while there are many more cultures that could be studied. While the authors filtered for unethical or harmful content in the word groups, it's very unlikely, but there is a possibility that we might have overlooked some.

References

Deepl translator. https://www.deepl.com/en/translator.

Google translate. https://translate.google.com/?sl=auto&tl=en&op=translate. Accessed on June 13, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Tian Liang, Zhiwei He, Jen tse Huang, Wenxuan Wang, Wenxiang Jiao, Rui Wang, Yujiu Yang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Leveraging word guessing games to assess the intelligence of large language models. *Preprint*, arXiv:2310.20499.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *Preprint*, arXiv:2404.12464.

Omar Shaikh, Caleb Ziems, William Held, Aryan J Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. *arXiv preprint arXiv:2306.02475*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. *arXiv preprint arXiv:2205.12247*.

A Related Work

Evaluating cultural bias in LLMs through QA A significant amount of work has been done in culture-based dataset creation. While varying in their sources, whether utilizing a pipeline (Shi et al., 2024) or manual creation utilizing existing sources or text generation (Durmus et al., 2023; Rao et al., 2024; Yin et al., 2022), these datasets can be utilized to evaluate cultural bias in LLMs. The datasets contain QA-style information, pairing cultural-based questions with correct answers depending on cultural knowledge. This includes objective information regarding cultural customs (Shi et al., 2024; Rao et al., 2024), information on a population's opinions (Durmus et al., 2023), and integrating geographic information (Yin et al., 2022). However, by evaluating information recall, there remains a gap in evaluating cultural-based reasoning, as knowledge may not necessarily omit the presence of bias. Our work introduces a dataset and associated task focusing on reasoning utilizing cultural concepts, rather than purely knowledge of cultural concepts themselves.

The usage of game-like assessments Much work has been done in the space of utilizing games as evaluation methods, assessing the reasoning abilities of an LLM. Previous work has utilized games either to benchmark reasoning within models (Liang et al., 2023) or as a means to assess the efficacy of prediction paradigms (Shaikh et al., 2023), among other varying uses. Comparatively, however, little work has been done in using games as a cultural assessment of LLMs. Our work strives to fill this gap, using a game as a vehicle to assess reasoning, and basing the game's content on aspects of culture.

B Dataset

B.1 Word Group Collection Details

Word Group Seed Data Details Each of the word groupings contained four words and an associated topic connecting each group. The Latinx/Hispanic culture groupings were collected from 7 individuals with a North/Central-American background, being composed of 60% colloquial (ex. Mexican Directors) and 40% non-colloquial (ex. Aquatic Animals) group topics. The Chinese culture groupings were collected from 5 individuals with a Chinese background, being composed of 44% colloquial (ex. The Five Classics) and 56% non-colloquial (ex. Parts of a Tree) group topics. Each of the groups was cross-verified by the authors to ensure adherence to game rules and representation of each culture.

Word Group Translation The translation is done by the authors manually (from being native speakers of each language used) and through the usage of online translation tools Google Translate (goo) and DeepL (dee).

B.2 Game Creation Details

200 2-, 3-, and 4-group games were sampled from each of the word group datasets, creating 12 datasets that were equally split into dev and test subsets. This dataset was manually checked by the authors to ensure that there is only one correct solution.

Each word group dataset was divided into subsets by the three group primary tags (described in 2.3), with each subset then being used to sample 100 2-, 3-, and 4-group games. This created 36 additional game datasets of 100 games each for use in tag-specific evaluation that were equally split into dev and test subsets.

C Determining Topic Achieved

The threshold value of 0.3 was selected through experimentation. The basis of the Topic Achieved score was evaluated using both BERTScore and FastText embeddings-based cosine similarity as measures of similarity between topics. 7 Different threshold values were used for score generation (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, and 0.7). For the 100-game 2-group C dataset, we had human annotations of topic achieved scores. Human annotators were provided the same rules for calculating the Topic Achieved score but were asked to base it on their own perception of similarity. As a means of selecting the metric most similar to human judgment, we calculated the inner-annotator agreement utilizing Randolph's Kappa between the human annotations and each of the methods used for calculation. From this, we found that a FastText embeddings-based approach with a similarity threshold of 0.3 performed best, attaining a Randolph's Kappa score of 0.55. A BERTScore-based approach with a similarity threshold of 0.5 had the second-best score of 0.53.

D Prompts

From these prompts, "#*#" is replaced with the total number of words in the pool, $*$ is replaced with the total number of groups in the game, and is replaced with the provided pool of words for a given game. A new prompt is generated for each new game, with each game's metrics being used to fill these spaces in the prompts.

English Prompt for 4-group games:
I am going to give you a pool of #*# words. These #*# words can be separated into $*$ equal groups of 4 words linked under some category. I want you to tell me the four groups and what category you think connects them. Here is an example: Given the pool ['Water', 'Fire', 'Sad', 'Wind', 'Happy', 'Earth', 'Angry', 'Surprised'], you would output: <Natural Elements>: ['Water', 'Fire', 'Earth', 'Wind'] , <Emotions>: ['Happy','Sad','Angry','Surprised']. Now, given the pool: . The answer must be $*$ groups, each of them containing 4 words and defined by one category, and the output format must be the same as the example. Give me the answer immediately.

English Prompt for 3-group games:
I am going to give you a pool of #*# words. These #*# words can be separated into $*$ equal groups of 4 words linked under some category. I want you to tell me the four groups and what category you think connects them. Here is an example: Given the pool ['Water', 'Happiness', 'Fire', 'Earth', 'Mercury', 'Surprise', 'Wind', 'Sadness', 'Venus', 'Pluto', 'Angry', 'Mars'], you would output: <Natural Elements>: ['Water', 'Fire', 'Earth', 'Wind'], <Emotions>: ['Happiness', 'Sadness', 'Angry', 'Surprise'], <Planets>: ['Mercury', 'Venus', 'Pluto', 'Mars']. Now, given the pool: . The answer must be $*$ groups, each of them containing 4 words and defined by one category, and the output format must be the same as the example. Give me the answer immediately.

English Prompt for 2-group games:
I am going to give you a pool of #*# words. These #*# words can be separated into $*$ equal groups of 4 words linked under some category. I want you to tell me the four groups and what category you think connects them. Here is an example: Given the pool ['Mile', 'League', 'Jazz', 'Heat', 'Yard', 'Cabaret', 'Carousel', 'Nets', 'Gobble', 'Scarf', 'Foot', 'Bucks', 'Chow', 'Wolf', 'Cats', 'Chicago'], you would output: <NBA TEAMS>: ['Bucks', 'Heat', 'Jazz', 'Nets'], <UNITS OF LENGTH>: ['Foot', 'League', 'Mile', 'Yard'], <Synonyms For Eat>: ['Chow', 'Gobble', 'Scarf', 'Wolf'], <Musicals Beginning With 'C'>: ['Cabaret', 'Carousel', 'Cats', 'Chicago']. Now, given the pool: . The answer must be $*$ groups, each of them containing 4 words and defined by one category, and the output format must be the same as the example. Give me the answer immediately.

E LLM Parameter Settings

For the efficiency and accuracy of our evaluation process, we fixed the same model parameters for both of LLama2-7B-chat and Mistral-7B-Instruct-v0.2 during different game evaluations.

For LLama2-7B-chat, Mistral-7B-Instruct-v0.2:

- do_sample: *True*
- top_k: *50*
- top_p: *0.9*
- temperature: *0.6*
- num_return_sequences: *1*
- max_length: *512*