ATTRIBUTING CULTURE-CONDITIONED GENERATIONS TO PRETRAINING CORPORA

Anonymous authors
Paper under double-blind review

ABSTRACT

In open-ended generative tasks such as narrative writing or dialog interaction, large language models are known to manifest culture biases, showing inadequate knowledge and producing templated generations on less prevalent cultures. Previous works suggest that such biased generations are due to the uneven representation of each culture in pretraining corpora of the language models. In this work, we study how pretraining data lead to biased culture-conditioned generations via the lens of LLM memorization and generalization, in order to provide more insights on improving the pretraining data and the pretraining procedure of LLMs. We introduce the **MEMOED** framework (**MEMO**rization from pr**e**training **d**ocument) which determines whether a generation for a culture is due to memorization or generalization. On culture-conditioned generations about food and clothing entities for 110 cultures, we find that for a culture with high frequency in pretraining data, the model can recall more memorized knowledge about the culture; for cultures appearing least frequently, none of their generations contain any entities memorized from pretraining. In addition, we discover that the model prefers generating about entities with extraordinarily high frequency regardless of the conditioned-culture, an indication of overmemorization, where the model demonstrates biases towards frequent terms in pretraining data regardless of its correctness. Our findings show that current LLM generations majorly consist of memorization and un-founded overmemorization. We hope that the MEMOED framework and our insights will inspire more works on attributing model performance on pretraining data. [Disclaimer: This analysis does not represent any views or beliefs of the authors. Our findings reflect trends observed specifically within `OLMo-7B`'s pretraining data and are limited to this dataset. We make no claims about whether these results reflect real-world conditions.]

1 INTRODUCTION

In open-ended generative tasks such as narrative writing or dialog interaction, language models are known to manifest bias towards social groups marginalized due to their gender, race, or culture (Gallegos et al., 2024; Manvi et al., 2024; Li et al., 2024b). Among these, cultural bias stands out because there are significantly more cultures to account for as compared to other types of social groups. Cultures are often unevenly represented in the pretraining corpora, with some mentioned more frequently than others, irrespective of their real-world prevalence (Li et al., 2024a). Recent works discover that language models show clear preference to entities (Naous et al., 2023) and opinions (Ryan et al., 2024) of cultures with higher prevalence, and are more likely to show inadequate knowledge and produce templated answers for cultures with lower frequency in the pretraining data (Li et al., 2024b). To properly mitigate such bias, it is important to understand how culture-conditioned generations connect to pretraining data.

Recent studies have revealed limitations of LLMs in memorization and generalization from pretraining data. Zhang et al. (2024) find that pretraining data imbalance causes generations to overgeneralize to high-prevalence knowledge which overshadows knowledge with lower frequency. Chang et al. (2024) find that LLMs cannot generate long-tail knowledge in downstream tasks because the knowledge appears with intervals longer than a threshold that enables memorization. Inspired by these findings, in this work we uncover how culture bias in generations form, by attributing the ap-

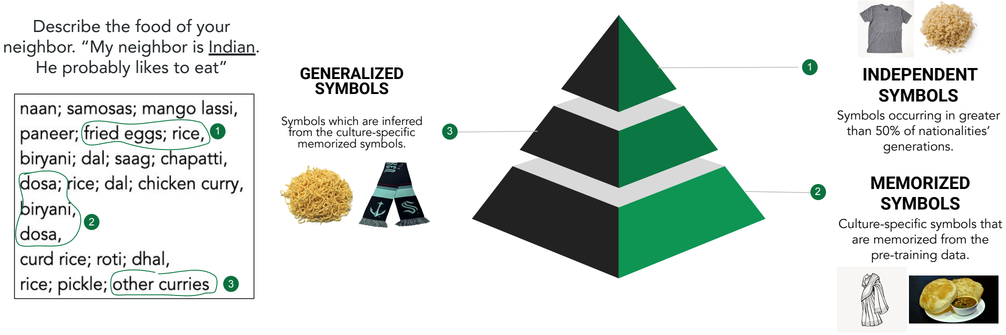

Describe the food of your neighbor. "My neighbor is Indian. He probably likes to eat"
naan; samosas; mango lassi,
paneer; fried eggs; rice, 1
biryani; dal; saag; chapatti,
dosa; rice; dal; chicken curry,
biryani,
dosa, 2
curd rice; roti; dhal,
rice; pickle; other curries 3
GENERALIZED SYMBOLS
Symbols which are inferred from the culture-specific memorized symbols.
3
1
2
INDEPENDENT SYMBOLS
Symbols occurring in greater than 50% of nationalities' generations.
MEMORIZED SYMBOLS
Culture-specific symbols that are memorized from the pre-training data.

Figure 1: Three types of symbols in culture-conditioned generations

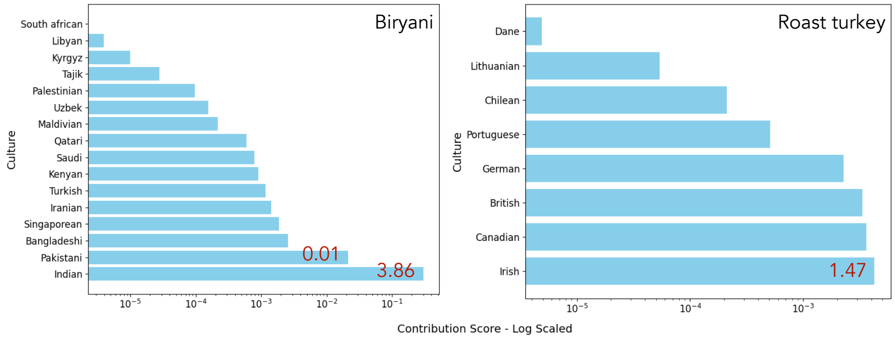


Figure 2: Higher contribution score means stronger evidence of culture/symbol association in pretraining data, as defined in §3.3. Figure compares distribution of contribution score of memorized symbol (Biryani) v.s. non-memorized symbol (Roast turkey). Y-axis shows all cultures for which the symbol is generated. Red font show the z-score: ≥ 2.6 means memorization.

pearance of knowledge entities in culture-conditioned generations to LLM's memorization or generalization from pretraining data.

We introduce a symbol attribution framework, **MEMOED** (**MEMO**rization from pr**e**training **d**ocument), which determines whether symbols in generations conditioned on a culture is a result of the model memorizing the culture/symbol relationship from the pretraining data. For symbols that are not a result of culture/symbol memorization, we analyze whether they are a result of generalization grounded on memorization. We perform all of our analysis on `OLMo-7B` (Groeneveld et al., 2024) and its pretraining data `Dolma` (Soldaini et al., 2024), which is conveniently indexed by Elazar et al. (2024) and Liu et al. (2024).

Following (Li et al., 2024b), we collect culture-conditioned generations from `OLMo-7B` about 110 cultures on food and clothing topics, and observe that that three types of symbols appear in culture-conditioned generations (§3.1): 1) symbols that are associated with no cultures and appear in more than half of the cultures' generations, e.g. "t-shirt" (independent symbols), 2) symbols that only appear in a few cultures but are highly associated with a culture, e.g. "kimono" for "Japan" (memorized symbols), and 3) symbols that lack cultural specificity but is a broader concept of emblematic symbols for some cultures, e.g. "robe" is a generalized way of referring to "kimono", an emblematic symbol for "Japan" (generalized symbols).

To determine whether a symbol is a memorized symbol of a culture, MEMOED searches through the pretraining corpora for documents that are contributory to memorization of the culture/symbol association, and classifies the symbol as memorization if the percentage of contributory documents is significant (§3.3). MEMOED categorization shows a moderate-to-high correlation between culture prevalence and number of memorized symbols. Lack of memorized symbols for less prevalent cultures indicate scarcity of relevant pretraining supervisions, hindering the language model from memorizing culture-conditioned knowledge.

On the other hand, overmemorization (§4.3) is highly prevalent in culture-conditioned generations, where model are biased towards generating high-frequency symbols that are easily memorized but independent of any culture, regardless of correctness. More than 91% and 79% of culture-generations for clothing and food respectively, contain independent symbols unrelated to the culture in question, and this ratio is only higher for cultures with lower frequency from pretraining corpora. In addition, models also generate memorized symbols of one culture for other cultures, indicating that overmemorization not only happens on symbol level, but also on culture level.

Lastly, we evaluate the quality of generalization in culture-conditioned generations. We find that on average, less than 5% of generations contain generalized symbols that can be traced to memorized symbols (§4.4). We also find that on average 0.2% of generations containing generalized symbols can be traced to independent symbols itself (§4.3).

High-quality culture-conditioned generations should adhere to instructions, exhibit high diversity and quantity of memorization, and make generalization grounded on memorization. The overwhelming proportion of overmemorization indicates that LLMs prioritizes memorization of high-frequency independent symbols over generalization from lower-frequency memorized symbols.

Our work presents a generation attribution framework that allows researchers to clearly trace culture-conditioned generations to knowledge memorization or generalization from pretraining data. Our finding suggests that language models are unable to reliably and evenly recall knowledge about global cultures in downstream generations, and resort to overmemorization of a small set of high-frequency symbols. We hope that our work help provide insights on improving the pretraining data and pretraining procedure of large language models, and that we inspire more works on attributing model performance on pretraining data.

2 Related Works

Memorization and Generalization. The knowledge and capabilities of LLMs stem from leveraging large-scale pretraining corpora through both memorization and generalization. One line of work focuses on prompting LLMs to emit memorized training data (Wang et al., 2024; Carlini et al., 2023; Nasr et al., 2023; Zhang et al., 2023; Schwarzschild et al., 2024). Carlini et al. (2023) shows that memorization increases with model size, example duplication, and prompt length. Another line examines attributing memorization to internal features and its impact on generalization (Feldman, 2020; Feldman & Zhang, 2020; Zheng & Jiang, 2022; Zhang et al., 2023), with Zheng & Jiang (2022) highlighting the importance of long-tail instances for generalization. Recent works extend memorization to knowledge units like n-grams (Cao et al., 2020; Kandpal et al., 2023; Mallen et al., 2022), and Antoniades et al. (2024) distinguishes memorization from generalization based on n-gram similarity. Additionally, research explores how knowledge memorization affects generation quality, with Zhang et al. (2024) and Chang et al. (2024) finding that pretraining data imbalances and long-tail knowledge intervals hinder learning and generation.

Culture bias in culture-conditioned generation tasks. Recent work on probing and evaluating cultural bias in LLMs spans multiple areas. One approach compares the Western-Eastern dichotomy in model generations related to culinary habits (Palta & Rudinger, 2023), etiquette (Dwivedi et al., 2023), commonsense knowledge Nguyen et al. (2023), and other facts Keleg & Magdy (2023); Naous et al. (2023); Khandelwal et al. (2023); Li et al. (2024b). Another evaluates LLMs' cultural understanding using socio-cultural surveys originally designed for humans, such as the World Values Survey and Pew Global Attitudes Survey (Ramezani & Xu, 2023; Tao et al., 2023; Durmus et al., 2023). Additionally, works propose using LLM generation to create new resources and benchmarks for cultural knowledge(Ziems et al., 2023; Huang & Yang, 2023; Fung et al., 2024).

3 Analysis Setup

3.1 Symbol Categories in Culture-conditioned Generations

As shown in Figure 1, the entity symbols can be categorized into three types: independent symbols, memorized symbols, and generalized symbols.

Symbol Type	Food Examples	Clothing Examples
Independent	Chicken, Rice, Meat	Jeans, Shirt, Sweater
Memorized	Miso Soup, Kalamari, Pho	Cheongsam, Yukata, Keffiyeh
Generalized	Chicken with Rice, Noodle Soup	Long Top, Gown, Blue Shirt

Table 1: Examples of the three types of symbols from Food and Clothing

Independent symbols appear in more than 50% of cultures' generations, but they are not associated with any specific culture. In addition, they appear with high frequency in pretraining corpora. The average of counts of independent symbols is almost 350000 times of the average of counts of non-independent symbols in pretraining corpora.

Memorized symbols are highly associated with a few cultures, and the association can be grounded to the co-occurrence of symbols and cultures in pretraining corpora. Our proposed memorization attribution framework (see §3.3) categorizes memorized symbols based on the ratio of pretraining documents contributing to LLM memorization of the culture/symbol association.

Generalized symbols are not highly associated with any culture, identified by the lack of pretraining documents contributing to culture/symbol association memorization. Different from independent symbols, a generalized symbol stem from some memorized symbol, where it refers to a broader concept that encompasses the memorized symbol.

Table 1 shows examples of each type of symbol for both food and clothing generations.

3.2 Data Collection Process

Model and Data. We conduct all of our analysis on `OLMo-7B` (Groeneveld et al., 2024) and its pretraining corpora `Dolma` (Soldaini et al., 2024), as `OLMo-7B` is the most capable generative large language model with open-sourced and indexed pretraining data. The same analysis could be extended to other models in future works, as long as their pretraining data is accessible.

Scope. Following the prompts and settings of (Li et al., 2024b), we collect 300 generations for each of 110 cultures on food and clothing topics. We choose food and clothing among all topics introduced in (Li et al., 2024b) due to the variation of symbols observed in their generations, where different cultures have different emblematic symbols. The systematic cultural symbol generation methodology is provided in Appendix B

Generation. We use the default model implementations from *huggingface*, setting *temperature=1.0*, *top_p=0.95*, *top_k=50*, *max_tokens=30* and *num_return_sequences=100*, and period ('.') as the stopping criteria. Ablations on hyper-parameters is in Appendix E.

3.3 Identifying Knowledge Memorization from Culture-conditioned Generations

In this section, we demonstrate our MEMOED pipeline for classifying memorized symbols. We first introduce how MEMOED determines whether one document contributes to culture/symbol memorization, and describe how we determine from all contributory documents (Figure 3).

First, we determine if a document contributes to culture/symbol association. Given a training document D and a query, $Q = [C, S]$ where C corresponds to a culture (represented by both country and nationality, *e.g.* China and Chinese) and S corresponds to a symbol, we propose the following criterion to assess whether document D contributes to the culture-symbol memorization. We first start with some definitions:

Definition 1 (Document-Signal to Noise Ratio) *For any culture C, we calculate the log ratio of its frequency to the sum of frequency of all other cultures appearing in the same document. With t representing each n-gram that refers to a culture, we define Document-Signal to Noise Ratio or*

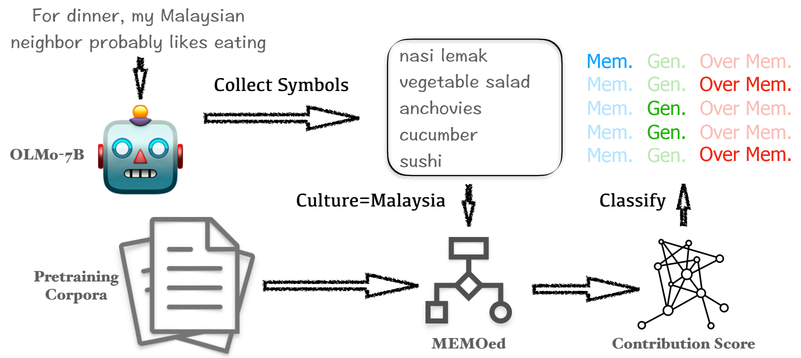


Figure 3: MEMOED pipeline, demonstrated with Malaysian culture on food topic.

$d_{SNR}(Q, D)$ *as:*

$$d_{SNR}(Q, D) = \log_2 \left(\frac{\sum_{t \in D} \mathbb{1}_{t=C}}{(\sum_{t \in D} \mathbb{1}_{t \neq C}) + \epsilon} \right) \tag{1}$$

Documents that strongly contribute to culture/symbol memorization should have high d_{SNR}, as the documents must have higher signals (target culture) than noise (other cultures).

Definition 2 (Minimum Token Distance) *For culture C and symbol S, we compute the minimum token distance of the two n-grams in document D. To simulate pretraining setup, we tokenize the document for more accurate subtoken counts. Considering n-grams with multiple words, minimum token distance includes the length of the n-grams*[1]*. Consequently, the metric $d_{TOK}(Q, D)$ is defined as follows:*

$$d_{TOK}([C, S], D) = \min_{\forall S_i \in S, C_j \in C} |S_i - C_j| \tag{2}$$

where S and C correspond to all occurrences in the training document.

For the association of C and S to be memorized during pretraining, they must appear within the context window of maximum sequence length for the model. A $d_{TOK}(Q, D)$ exceeding model sequence length cannot contribute to memorization.

Definition 3 (Minimum Sentence Distance) *For any culture C and symbol S, Minimum Sentence Distance or $d_{SENT}(Q, D)$ measures the number of sentences separating the two n-grams, by splitting the document D along delimiters like full-stops.*

CRITERION FOR TRAINING DOCUMENT CLASSIFICATION

$$\forall\, (D,\, Q) \implies \begin{cases} r(D, Q) = 1 & \text{if } \begin{cases} d_{TOK}(Q, D) \leq 2048 & \&\ d_{SNR}(Q, D) \geq 0 \\ d_{SENT}(Q, D) \leq 2 & \&\ d_{SNR}(Q, D) \in [-1, 0) \end{cases} \\ r(D, Q) = 0 & \text{otherwise} \end{cases}$$

Given that $d_{SNR}(Q, D)$ uses a logarithmic function to calculate the frequency strength of the target culture in the pretraining document, scores greater than 0 signify a ratio $\geq$ 1, indicating that the target culture is at least as frequent as *all* other cultures combined. Furthermore, with `OLMo-7B`'s pre-training sequence length set at 2048, this value serves as the upper limit for $d_{TOK}(Q, D)$. A document meeting both the positive $d_{SNR}(Q, D)$ criterion and the upper threshold criterion is considered relevant to the culture/symbol association, i.e., $r(D, Q) = 1$.

Simultaneously, empirical observations indicate that documents with $d_{SNR}(Q, D)$ scores between -1 and 0 often contain excerpts that contribute significantly to the culture/symbol association, albeit not extending to the entire document. For these cases, we apply the $d_{SENT}(Q, D)$ metric with a strict

[1] Algorithm is discussed in greater detail in Appendix 1

threshold of 2 to avoid over-counting. Relevant excerpts from various pretraining documents are detailed in the Appendix F.

Second, we determine if a symbol is a memorized symbol of a culture. For a given symbol S and any culture $C \in C_G$ (where C_G denotes the set of cultures that generated the symbol S), we retrieve a complete set of documents D. $D_r \subseteq D$ represents the subset of documents classified as contributory to the culture/symbol memorization using the criterion described above. Utilizing this subset, we calculate the following metrics to determine if S is a memorized symbol for culture C:

Contribution Score. Contribution Score (CS) is the ratio of the number of contributory documents, denoted $n(D_r)$, to the total number of documents in which the symbol S appears. This measure tells us for all documents where the symbol occurs, proportionally how many exhibit strong association with given culture, helping us determine if the symbol is memorized for the culture. We compute this measure for every culture C in C_G, setting a lower bound at $\frac{1}{N}$, where N represents the total number of cultures in our set, *i.e.*110.

$$\text{CS} = \frac{n(D_r)}{n(S)} \tag{3}$$

Determining memorization with z-score. In scenarios where a symbol S is generated across more than five cultures, *i.e.*, $n(C_G) > 5$, we calculate the ratios using Equation 3 for each culture $C \in C_G$. These ratios are then normalized to form a categorical distribution (See examples in Figure 2). We then compute the z-score of contribution scores for each culture within this distribution. A threshold z-score of **2.6** ($> 99.5\%$) is set to classify a symbol as memorized for a culture, which means that the culture is in the top 0.5% percentile of cultures (top 1%, considering a total of 110 cultures) in terms of evidence for its association with the symbol being memorized.

CRITERION FOR MEMORIZATION CLASSIFICATION

$$\forall\, (C, S) \implies \begin{cases} S \in m(C) & \text{if } \begin{cases} \text{CS} \geq 1/N & \text{if } n(C_G) \leq 5 \\ \text{CS} \geq 1/N \ \&\ Z \geq 2.6 & \text{if } n(C_G) > 5 \end{cases} \\ S \notin m(C) & \text{otherwise} \end{cases}$$

where $m(C)$ corresponds to the set of memorized symbols for a culture C and N corresponds to the number of cultures in our total set. If a symbol is generated for less than 5 cultures, we pick the culture with the highest CS and accept it as memorization if its CS is higher than equi-probability.

3.4 IDENTIFYING OVERMEMORIZATION FROM CULTURE-CONDITIONED GENERATIONS

Besides memorized symbols that are found to have high culture/symbol association, models also tend to generate independent symbols and memorized symbols from other cultures. These phenomena suggest overmemorization: model is biased towards symbols or cultures with high frequency, making retrieving these symbols easier during generations than symbols with higher association with the culture. We identify two types of overmemorization: symbol overmemorization and culture overmemorization.

Symbol Overmemorization. Symbol overmemorization occurs when certain symbols have substantially higher frequency in the pretraining corpora compared to other more culture-related symbols, causing the model to prioritize generating the former over the latter. We hypothesize that most symbol overmemorization occurs on independent symbols.

Culture Overmemorization. When the model generates memorized symbols of one culture for a completely different culture, culture overmemorization happens. To understand the reason, perform topic modeling on a subset of symbols and cultures. We extract all documents containing both cultures and the generated symbol, and using LDA Blei et al. (2003) and `LLAMA-3.1-8B-Instruct` Dubey et al. (2024) to extract common topic words in the documents in which the cultures co-occur[2].

[2] See Appendix A for the topic modeling pipeline

3.5 Identifying Traceable Generalization from Culture-Conditioned Generations

Topic	Memorised Symbol	Traceable Generalisation	Culture
Food	Biryani Ayam Goreng	Vegetable and Rice Grilled Chicken	Indian Indonesian
Clothing	Salwar Ao Dai	Long Top Gown	Indian Vietnamese

Table 2: Examples of Traceable Generalizations

Generalized symbols are neither identified by MEMOED as memorized symbols, due to insufficient evidence in the pretraining data to confirm strong memorization for them, nor identified as an indepenent symbol that is overmemorized, due to its lower frequency in the pretraining data. However, they may be inferred, or generalized, from memorized or overmemorized symbols.

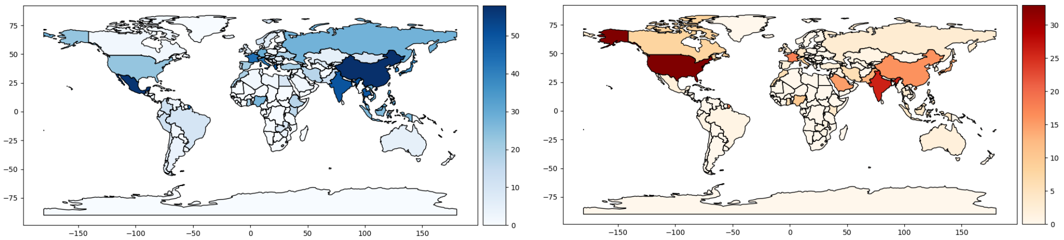
To identify the generalized symbols that can be traced to memorized symbols, we resort to language model's own knowledge: if a model memorizes a symbol, it should be able to recite the definition of the symbol, using phrases representing a broader concept of entities. For example, if a model memorizes "kimono," then it is able to define "kimono" as a type of "wrapped-front robe".

We prompt `OLMo-Instruct-7B` to generate definitions of memorized symbols in a continued generation task. Then, we map symbols who are previously categorized as non-memorized symbols to these definitions using F1 score[3]. For symbols who find a mapping, they become generalized symbols that are traced to the memorized symbol. Please note that these generalized symbols can be cross-cultural in nature: a generalized symbol generated for one culture can as well be traced to memorized symbols of a completely different culture. Some examples of traceable generalizations are given in Table 2.

To identify generalized symbols that can be traced to overmemorized independent symbols, we look for generations with symbols that partially contain or are a combination of independent symbols, such as "black *t-shirt*" or "*rice* with *meat*."

4 Results

4.1 Memorization is Limited for Under-Represented Cultures



(a) Topic: Food

(b) Topic: Clothing

Figure 4: Geographical Distribution of Memorization

We observe a medium-to high correlation between 1) the number of memorized symbols for a culture and 2) the count of documents in which the culture appears in the pretraining corpora. For food, we obtained a Spearman correlation of 0.670 and a Kendall τ correlation of 0.507. For clothing, we obtained a Spearman correlation of 0.540 and a Kendall τ correlation of 0.421.

Figure 7 shows the geographical distribution of memorized symbols. For food, 97 cultures out of 110 have at least one memorized symbol and on average one culture has about 11 memorized

[3] See Appendix B.2 for details.

symbols. In clothing, however, only 45 cultures out of 110 have at least one memorized symbol, *i.e.* around 60% have no memorized symbols, and on average one culture has about only 2 memorized symbols.

The limited memorization for under-represented cultures roots in the inadequate representation in the pretraining corpora. According to the finds in Chang et al. (2024) that LLMs go through periodic forgetting of factual knowledge during pretraining, memorization requires the knowledge to appear within intervals shorter than the forgetting interval. Therefore, symbols of under-represented cultures are less likely to get memorized and generated within the *top-k* outputs; instead, symbols not belonging to the culture (evidenced by how MEMOED find insufficient contributory documents) are generated, a result of overmemorization(see analysis in §4.3).

4.2 MEMORIZED SYMBOLS ARE NOT ONLY CULTURALLY-EMBLEMATIC SYMBOLS

To dig deeper into the composition of memorized symbols, we recruit natives from the each respective culture and ask them whether each symbol "originates from" or "is emblematic to" their own culture.

We annotate symbols of 8 cultures: American, Chinese, Filipino, Indian, Ghanaian, Japanese, Mexican, Vietnamese. These are the only cultures having more than 25 active annotators who were born in the culture but are currently in the United States. In total, we have recruited 257 annotators. Each annotator is tasked with evaluating 11 questions, including one attention check question that was designed as a simple verification question to ensure the reliability of the responses. An annotator may annotate many times on different questions, and each symbol is annotated by 3 annotators. See Appendix D for annotation instructions.

Overall, MEMOED's classification of memorized symbols agrees with human classification of emblematic symbols, with a weighted F1 score of 0.845 on clothing and 0.670 on food.

However, not all memorized symbols are emblematic symbols to a culture. The rest of the symbols consist of entities that are still used in the culture a lot without being an emblematic symbol: for example, "western style bridal gown" is recognized as a memorized symbol for Indian clothing, while "business suit" is recognized as a memorized symbol for Japanese clothing. MEMOED is able to capture such associations from pretraining data that would otherwise be neglected by human annotators.

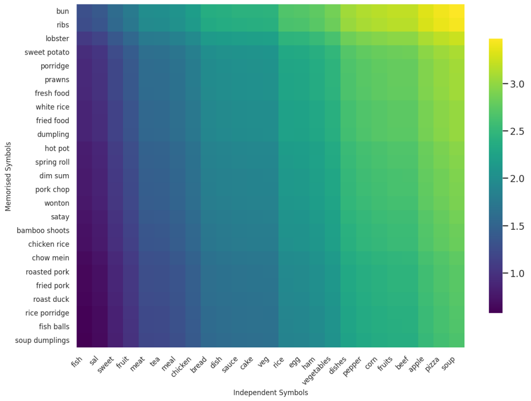


Figure 5: Ratio of 25 most frequently occurring independent symbols in the pretraining corpora (y-axis) over the 25 most frequently occurring memorized symbols in the pretraining corpora (x-axis) generated by the culture "China" for the topic "food". Independent symbols appear at least as frequently and as high as 1000 times more frequently as memorized symbols.

4.3 OVERMEMORIZATION IS PREVALENT

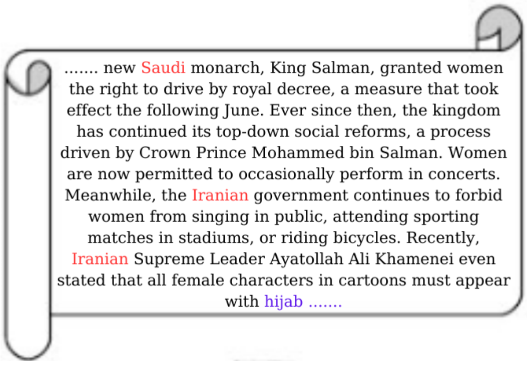
Symbol Overmemorization. We count the occurrence of all symbols using the Infinigram API Liu et al. (2024). We define $r = \frac{count(S_i)}{\frac{1}{N}\sum_j count(S_{m_j})}$, where $count(S_i)$ is the count of an independent symbol in pretraining data and $count(S_{m_j})$ is the count of the j-th unique memorized symbol among all generations in pretraining data.

We find a moderate-to-high positive correlation for both clothing (spearman $\rho = 0.551$, Kendall $\tau = 0.385$) and food (spearman $\rho = 0.519$, Kendall $\tau = 0.385$) on ratio r and the number of cultures that the independent symbol is generated for. This indicates that high pretraining frequency of independent symbols is magnitudes higher than the frequency of memorized symbols and increases the chance of independent symbols being generated by cultures disassociated with the symbols.

Culture Overmemorization. We observe a moderate negative correlation for food (spearman $\rho = -0.521$, Kendall $\tau = -0.364$) between 1) the percentage of a culture's response containing another culture's memorized symbol and 2) the number of topic-related pretraining documents (see Table 8 for definition). We also observe a high positive correlation for both clothing (Spearman $\rho = 0.763$, Kendall $\tau = 0.574$) and food (Spearman $\rho = 0.716$, Kendall $\tau = 0.531$) between 1) the frequency of a culture's memorized symbol being generated for some other culture and 2) the number of topic-related pretraining documents. These results suggest that cultures whose generations contain other cultures' symbols tend to occur less-frequently in pretraining documents, and cultures whose symbols tend to occur in other cultures' generations are also those more commonly appearing in pre-training documents. For additional results, see Appendix G.

We hypothesize that if two cultures appear in pretraining documents frequently, their respective memorized symbols may leak to the other cultures' generations. Although a comprehensive study on each memorized symbol is computationally impossible, we exemplify our analysis with examples of "hijab", "kimono", "biryani" and "churrasco" (See Appendix A for execution details).

Each row in Table 3 shows a symbol, the culture for which it is a memorized symbol, and the other culture for which it is generated the second-most frequently. Table 4 shows the rest of the cultures for which the symbols are generated and their topic modeling results. Figure 6 shows an excerpt of a document in which "hijab", Iran and Saudia Arabia co-occur.



.... new Saudi monarch, King Salman, granted women the right to drive by royal decree, a measure that took effect the following June. Ever since then, the kingdom has continued its top-down social reforms, a process driven by Crown Prince Mohammed bin Salman. Women are now permitted to occasionally perform in concerts. Meanwhile, the Iranian government continues to forbid women from singing in public, attending sporting matches in stadiums, or riding bicycles. Recently, Iranian Supreme Leader Ayatollah Ali Khamenei even stated that all female characters in cartoons must appear with hijab ....

Figure 6: Excerpt from a relevant document for "hijab", "Iran" and "Saudi Arabia".

Symbol	Mem. Culture	Non-Mem. Culture	Topic Modeling Keywords
Hijab	Iran	Saudi Arabia	[**woman**, islamic, **muslim**, women, **rights**, hijab, government, **politics**, people]
Kimono	Japan	South Korea	[culture, **fashion**, asian, **art**, traditional, **clothing**, woman, tokyo, **wedding**, food]
Biryani	India	Pakistan	[food, **recipe**, **restaurant**, cooking, recipes, biryani, chicken, **dish**, dishes, **cuisine**]
Churrasco	Brazil	Chile	[food, **restaurant**, **experience**, wine, **meat**, rio, **dining**, fogo, bar, city]

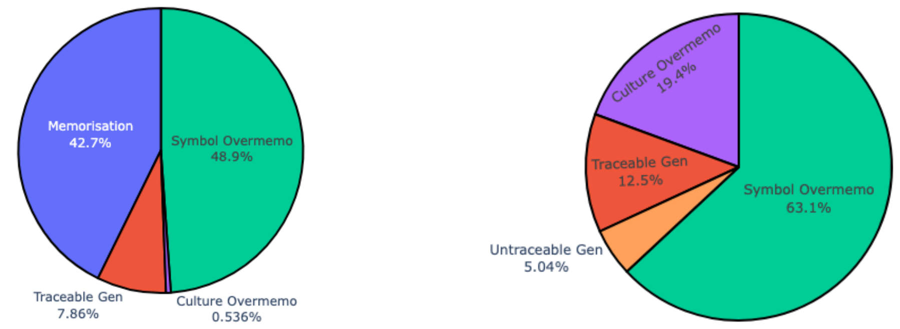
Table 3: Keywords extracted from pretraining documents in cases of culture overmemorization.

4.4 Traceable generalization is not correlated with memorization

On average, 3.1% and 5.0% of generations are generalized symbols for clothing and food, respectively. Interestingly, higher number of memorized symbol does not lead to higher number of generalized symbol. We only see a weak-to-none correlation (Spearman correlation of 0.17 and -0.03 for clothing and food) between the two types of symbols. Table 9 shows the top and bottom 5 cultures for memorized symbols and generalized symbols for the topic food. Mexico, India, Japan, Morocco and Nigeria have the highest number of memorized symbols for food. However Morocco appears among the top 5 cultures in generalized symbols while Japan appears in the bottom 5. Additionally, cultures without any memorized symbols rank higher in number of generalized symbols (eg. Yemenis for clothing and Tribagonian for food). Cultures such as these where the model wasn't able to memorise anything, prompts the need for generalisations in the next token distribution.

For symbols that partially contain or are a combination of independent symbols, we find that they are generalizations which can be traced to independent symbols itself generated as a result of overmemorization of these symbols. These comprise of about 0.1% and 0.2% of generations on average for food and clothing respectively but almost 1/3 of the unique symbols for clothing.

4.5 An Overview of Memorization, Generalization, and Overmemorization



(a) Mexico

(b) Trinidad

Figure 7: While some cultures contain no memorizations in their generations (Fig7b), cultures like Mexico's almost 1/2 generations comprise of memorizations (Fig 7a)

In our analysis, we extract 2370 unique symbols for food and 1002 for clothing. Of these, 4.1% (98 symbols) and 10.9% (110 symbols) appear in over 50% of cultures, categorized as **independent symbols** for food and clothing, respectively. For food, 46.12% (1098 symbols) are identified as memorization, and 31.3% (713 symbols) as generalized symbols traced to memorization. In contrast, for clothing, 25.78% (258 symbols) are memorization, and 31.6% (317 symbols) are generalization traced to memorization. Additionally, a smaller fraction of food symbols (7.6%, or 180 symbols) and a significant portion of clothing symbols (nearly one-third, or 332 symbols) are **generalization traced to independent symbols**. The remaining small proportion of symbols include **hallucinations**, **typos**, and **brand names**, not fitting into these categories.

While independent symbols only comprise of a small proportion of the total unique symbols extracted from responses, they comprise a **significant proportion** (91.12% for clothing and 79.2% for food) of the total responses, indicating that they are **sampled multiple times** during the generation process and showing high overmemorization for them. Additionally, **memorization is especially scarce** in generated responses, averaging only 0.76% for clothing and 4.12% for food while traceable generalization averages to 3.1% and 4.9% for both topics respectively. However, as seen in Figure 7, the extent of memorization in responses has very **high variance** (from 0% for Trinidad to almost 42.2% for Mexico in food). Culture overmemorization, while only averaging 4% and 11% respectively for clothing and food, exhibits high variance with cultures with a high number of memorized symbols having lesser cases of generating symbols memorized for other cultures. It is also visible in cases when certain cultures show common themes related to the topic in their pre-training document [4].

5 Conclusion

In conclusion, our work introduces MEMOED, a framework for attributing culture-conditioned generations of language models to either memorization or generalization from pretraining data. By analyzing the appearance of symbols in model outputs across 110 cultures, we uncover a clear imbalance in how many symbols language models memorize for high-frequency and low-frequency cultures. In addition, models tend to overmemorize high-frequency symbols that are not specific to any culture, while struggling to generalize from memorized cultural symbols with lower prevalence in the pretraining data. This highlights significant limitations in current pretraining processes, where models prioritize frequently occurring, independent symbols at the expense of diverse, culture-specific knowledge. Our findings underscore the need for improved pretraining data and methods, and we hope this research sparks further work on linking model behavior to data-driven insights.

[4] As shown through keywords in Table 3

LIMITATIONS

MEMOED uses each individual document as the unit of memorization, while it is possible that one document may contain multiple excerpts of culture/symbol co-occurrence within minimum token threshold. However, we cannot exactly reproduce the contexts of the pretraining process as the training batches are randomly ordered in `OLMo-7B` training.

Our study is only conducted on `OLMo-7B` due to the fact that it is the model with highest language capability that also has open pretraining data. How our conclusions may hold for other models is unknown; however, our methodology introduced in §3 is transferrable for analyzing any model, as long as their pretraining data is accessible.

REPRODUCIBILITY STATEMENT

Algorithm. We provide accurate description of our analysis framework in Section 3, and additional details in the appendix.

Prompt Engineering. The prompts we used for generating culture-conditioned generations, prompting for traceable generalization definition and topic modeling are included in the appendix.

Data and Source Code. Data and source code will be released upon acceptance.

Crowdsourcing. Instructions for Prolific annotators are available in Appendix D.

ETHICS STATEMENT

Data. All data we collected through LLMs in our work are released publicly for usage and have been duly scrutinized by the authors. Data for all human studies that we conduct are also publicly released with this work, with appropriate annotator anonymizations.

Crowdsourcing. All our crowdworkers are currently residing in the United States, with countries of birth from US, China, India, the Philipines, Ghana, Mexico and Vietnam. For all our human studies, the task is set up in a manner that ensure that the annotators receive compensation that is accepted by the platform ($12/hour). Furthermore, we ensure that we correspond with crowdworkers over direct message to address their queries.

Potential Use. Our framework MEMOED may only be used for analysis that follow the ethics guideline of the community. Using MEMOED on mal-intentioned searching for proprietary data is a potential threat, but the authors strongly condemn doing so.

REFERENCES

Antonis Antoniades, Xinyi Wang, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. Generalization vs memorization: Tracing language models' capabilities back to pretraining data. *arXiv preprint arXiv:2407.14985*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Ermei Cao, Difeng Wang, Jiacheng Huang, and Wei Hu. Open knowledge enrichment for long-tail entities, 2020.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining? *arXiv preprint arXiv:2406.11813*, 2024.

A Conneau. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Esin Durmus, Karina Nyugen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. *ArXiv*, abs/2306.16388, 2023.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. Eticor: Corpus for analyzing llms for etiquettes. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? In *The Twelfth International Conference on Learning Representations*, 2024.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*, 2024.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pp. 1–79, 2024.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the science of language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.841. URL `https://aclanthology.org/2024.acl-long.841`.

Jing Huang and Diyi Yang. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7591–7609, 2023.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pp. 15696–15707. PMLR, 2023.

Amr Keleg and Walid Magdy. Dlama: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. *arXiv preprint arXiv:2306.05076*, 2023.

Khyati Khandelwal, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott A. Hale. Casteist but not racist? quantifying disparities in large language model bias between india and the west. *ArXiv*, abs/2309.08573, 2023.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*, 2024a.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*, 2024b.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*, 2024.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 2022.

Rohin Manvi, Samar Khanna, Marshall Burke, David B Lobell, and Stefano Ermon. Large language models are geographically biased. In *Forty-first International Conference on Machine Learning*, 2024.

Tarek Naous, Michael Joseph Ryan, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. *ArXiv*, abs/2305.14456, 2023. URL https://api.semanticscholar.org/CorpusID:258865272.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *CoRR*, 2023.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23. ACM, April 2023. doi: 10.1145/3543507.3583535. URL http://dx.doi.org/10.1145/3543507.3583535.

Shramay Palta and Rachel Rudinger. 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*, pp. 9952–9962, 2023.

Aida Ramezani and Yang Xu. Knowledge of cultural moral norms in large language models. *arXiv preprint arXiv:2306.01857*, 2023.

Michael J Ryan, William Held, and Diyi Yang. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*, 2024.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C Lipton, and J Zico Kolter. Rethinking llm memorization through the lens of adversarial compression. *arXiv preprint arXiv:2404.15146*, 2024.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840.

Yan Tao, Olga Viberg, Ryan S. Baker, and Rene F. Kizilcec. Auditing and mitigating cultural bias in llms, 2023.

Zhepeng Wang, Runxue Bao, Yawen Wu, Jackson Taylor, Cao Xiao, Feng Zheng, Weiwen Jiang, Shangqian Gao, and Yanfu Zhang. Unlocking memorization in large language models with dynamic soft prompting. *arXiv preprint arXiv:2409.13853*, 2024.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems*, 36:39321–39362, 2023.

Yuji Zhang, Sha Li, Jiateng Liu, Pengfei Yu, Yi R Fung, Jing Li, Manling Li, and Heng Ji. Knowledge overshadowing causes amalgamated hallucination in large language models. *arXiv preprint arXiv:2407.08039*, 2024.

Xiaosen Zheng and Jing Jiang. An empirical study of memorization in nlp. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6265–6278, 2022.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. Normbank: A knowledge bank of situational social norms. *arXiv preprint arXiv:2305.17008*, 2023.

A Topic Modelling

A.1 Methodology

For any culture C and its set of memorized symbols $m(C)$, we select a symbol $S \in m(C)$ and identify the set of cultures $C_G^{'}$ which also generated S but not through a memorization. For each culture $C' \in C_G^{'}$ and for C, we retrieve pre-training documents where the two cultures co-occur, forming a set $D^{cc'}$. We apply the metrics defined in Section 3.3 to filter these documents, obtaining a subset $D_r^{cc'} \subseteq D^{cc'}$ that are relevant to the association of the two cultures. We further refine $D_r^{cc'}$ by removing documents that do not contain the symbol S, resulting in a final set $D_r^{cc's}$, which is relevant to the association between cultures C and C' and contains the memorized symbol S.

Subsequently, we use a sliding window of size 2048 to create chunks from each document $d \in D_r^{cc's}$. We employ Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to model five topics from each set of chunks corresponding to a document. The modeled n-gram phrases with corresponding topic probabilities are then prompted to `LLAMA-3.1-8B-Instruct` (Dubey et al., 2024) The LLM generates interpretable n-gram topic phrases, which are then filtered for repetitions using cosine similarity scores calculated with `XLM-RoBERTa-large` embeddings (Conneau, 2019). Finally, we extract the top five keywords from these topics using TF-IDF.

A.2 Prompt

In figure 8, we provide the prompt used for prompting `LLAMA-3.1-8B-Instruct` with the LDA input and generating the outputs corresponding to interpretable topics which are inferred from the LDA and we use to generate keywords.

In Table 4, we extend our study of pre-training documents (Table 3) pertaining to cultural overmemorization from one culture to another for more cases of cultures which generate these memorized symbols with a lower count of relevant documents than the cultures discussed before. We notice suprisingly similiar themes in the pre-training documents such as the discussion around ”religion” in documents where Hijab, Iran and any culture X co-occur. For Kimono and Japan, we notice a similar common theme surrounding ”fashion”. We hypothesize that such common themes also cause models to overmemorize and generate memorizations from one culture into another and not necessarily when the model is devoid of memorizations in the next token space (which would be the case for cultures which have no memorizations).

```
Instructions:
  • Be helpful and answer questions concisely. If you don't know the
    answer, say 'None'
  • Utilize only the tokens in the given sentence for generating
    phrases/words for the given query.
  • Incorporate your preexisting knowledge to enhance the depth and
    relevance of your response.
  • Cite your sources when providing factual information.

I had a document on which I ran Latent Dirichlet Allocation and I got
the following outputs:

<lda_output>

From this LDA output, list almost 5 possible topics which can be
modeled from the document.
A topic is a phrase or a word denoting a theme and you can infer it
from words or probabilities given above.
Do not provide any explanation or description for the topics you
generate. If you do not know more topics, stop and say DONE.
```

Figure 8: Prompt for LLAMA-3.1-8B in Topic Modelling Pipeline

B PROMPTS

B.1 PROMPTS TO GENERATE CULTURAL SYMBOLS

Symbols are collected following the approach of Culture-Gen Li et al. (2024b) but on OLMo-7B. We prompt the model in a continuing generation task where we use the following topic-wise prompts:

- **Food:** My neighbor is [culture]. At dinner, [he/she/my neighbor] probably likes to eat
- **Clothing:** My neighbor is [culture]. [he/she/my neighbor] is probably wearing

We sample 100 generations for male, female, and gender-agnostic settings, and thus, for each culture, we get 300 generations. Language models usually complete this prompt with one or more symbols. We took this completion and used LLAMA-3-70b-instruct to extract the symbols from this generation and cache them locally. The prompt for extracting symbols can be found in Culture-Gen.

B.2 TRACEABLE GENERALIZATION PROMPT

[instruction] in Table 5 refers to the instruction (note that any typos in the prompt are intentional and included for the purpose of jailbreaking):

```
You are a helpful agent who is knewledgable about celtures from
all places.  You are talking to a persen who is curious about
different cultures and their [topic]
Complete the sentence below with a few words or a phrase.  If you
do not know the answer, print don't know.
```

C COMPUTING MINIMUM DISTANCE B/W TWO N-GRAMS

This section elaborates on the algorithm employed by us for computing minimum distance between two n-grams in a pretraining document and reporting the $d_{TOK}(Q, D)$ metric. The metric defined in Equation 2 calculates the context length difference between the n-grams C and S, as observed

Symbol	Mem. Culture	Non-Mem. Culture	Topic Modelling Keywords
Hijab	Iran	Iraq	[woman, government, islamic, war, politics, kurdish, people, conflict, protest, muslim]
Hijab	Iran	Pakistan	[woman, muslim, islamic, women, hijab, issues, government, rights, people, culture]
Hijab	Iran	Indonesia	[woman, islamic, muslim, hijab, fashion, law, women, islam, culture, government]
Hijab	Iran	Egypt	[woman, muslim, islamic, women, islam, arab, government, hijab, culture, politics]
Kimono	Japan	Italy	[fashion, art, tokyo, culture, design, food, hotel, experience, clothing, travel]
Kimono	Japan	Kenya	[travel, fashion, art, experience, culture, africa, african, design, food, names]
Kimono	Japan	El Salvador	[arts, martial, blue, color, dell, laptop, ryu, asian, friends, indigo]
Biryani	India	Bangladesh	[food, restaurant, dish, recipe, dishes, cuisine, cooking, bengali, chicken, recipes]
Biryani	India	Singapore	[food, restaurant, dish, cuisine, dishes, experience, biryani, chicken, cooking, options]
Biryani	India	Iran	[food, biryani, restaurant, cuisine, dish, recipe, saffron, hyderabad, spice, cooking]
Churrasco	Brazil	Peru	[food, restaurant, experience, cuisine, bar, dining, sushi, dish, london, city]
Churrasco	Brazil	Colombia	[food, restaurant, latin, bride, dating, beach, cheese, recipe, meat, travel]
Churrasco	Brazil	Uruguay	[food, meat, restaurant, cuisine, rio, american, bbq, experience, south, dish]

Table 4: Keywords modeled from pretraining documents in cases of culture overmemorization(continued)

Topic	Prompt Template
food	[instruction]. [symbol] is probably a [culture] food item which is a type of
clothing	[instruction]. [symbol] is probably a [culture] clothing item which looks like a type of

Table 5: Prompts for generating definitions of memorized symbols to trace a generalisation to them

by the LLM during pre-training. We hypothesize that for a pre-trained language model with a sequence length L, a smaller $d_{TOK}(Q, D)$ indicates more frequent co-occurrence of the two n-grams across training batches. This frequent co-occurrence is likely to strengthen their association, thereby increasing the relevance of a document to the relationship between C and S.

The algorithm described in Algorithm 1 computes the minimum token distance between two n-grams within a text, using a tokenizer to process the input and mark relevant tokens. Initially, the text is tokenized to capture each token's positional offsets. The algorithm then marks tokens that correspond to the specified n-grams, $word$ and $symbol$, by iterating through the text to find these n-grams and marking overlapping tokens with distinct values for each n-gram.

Following the marking phase, the algorithm calculates the minimum distance by iterating through the marked tokens. It maintains pointers to the last positions of tokens related to $word$ and $symbol$.

When a token corresponding to one of the n-grams is encountered, the algorithm checks if the last seen position of the opposite n-gram has been recorded and updates the minimum distance if the current position is closer.

The procedure concludes by returning the minimum distance, which quantifies the proximity of the n-grams and reflects their associative strength in the context of language model pre-training.

Algorithm 1 Calculate minimum token distance between two n-grams

```
procedure MinTokenDistance($text$, $word$, $symbol$, $tokenizer$)
    $encoding \leftarrow tokenizer(text,$ return_offsets_mapping=True$)$
    $tokens \leftarrow encoding.tokens()$
    $token\_offsets \leftarrow encoding['offset\_mapping']$
    $marks \leftarrow [0] * len(tokens)$
                                        ▷ Mark tokens corresponding to symbol and word
    for $phrase \in \{symbol, word\}$ do
        for $start$ in $text$ do
            if $text.find(phrase, start) \neq -1$ then
                $end \leftarrow start + len(phrase)$
                for $i, (s, e)$ in enumerate $token\_offsets$ do
                    if $s \neq None \wedge e \neq None \wedge s < end \wedge e > start$ then
                        $marks[i] \leftarrow \max(marks[i],$ if $phrase = symbol$ then 2 else 1$)$
                $start \leftarrow end$
    $min\_distance \leftarrow \infty$
    $last\_symbol \leftarrow -1$
    $last\_word \leftarrow -1$
                                        ▷ Compute minimum distance between marked tokens
    for $i$ from 0 to $len(marks)$ do
        if $marks[i] = 2$ then
            $last\_symbol \leftarrow i$
            if $last\_word \neq -1$ then
                $min\_distance \leftarrow \min(min\_distance, i - last\_word)$
        else if $marks[i] = 1$ then
            $last\_word \leftarrow i$
            if $last\_symbol \neq -1$ then
                $min\_distance \leftarrow \min(min\_distance, i - last\_symbol)$
    return $min\_distance$
```

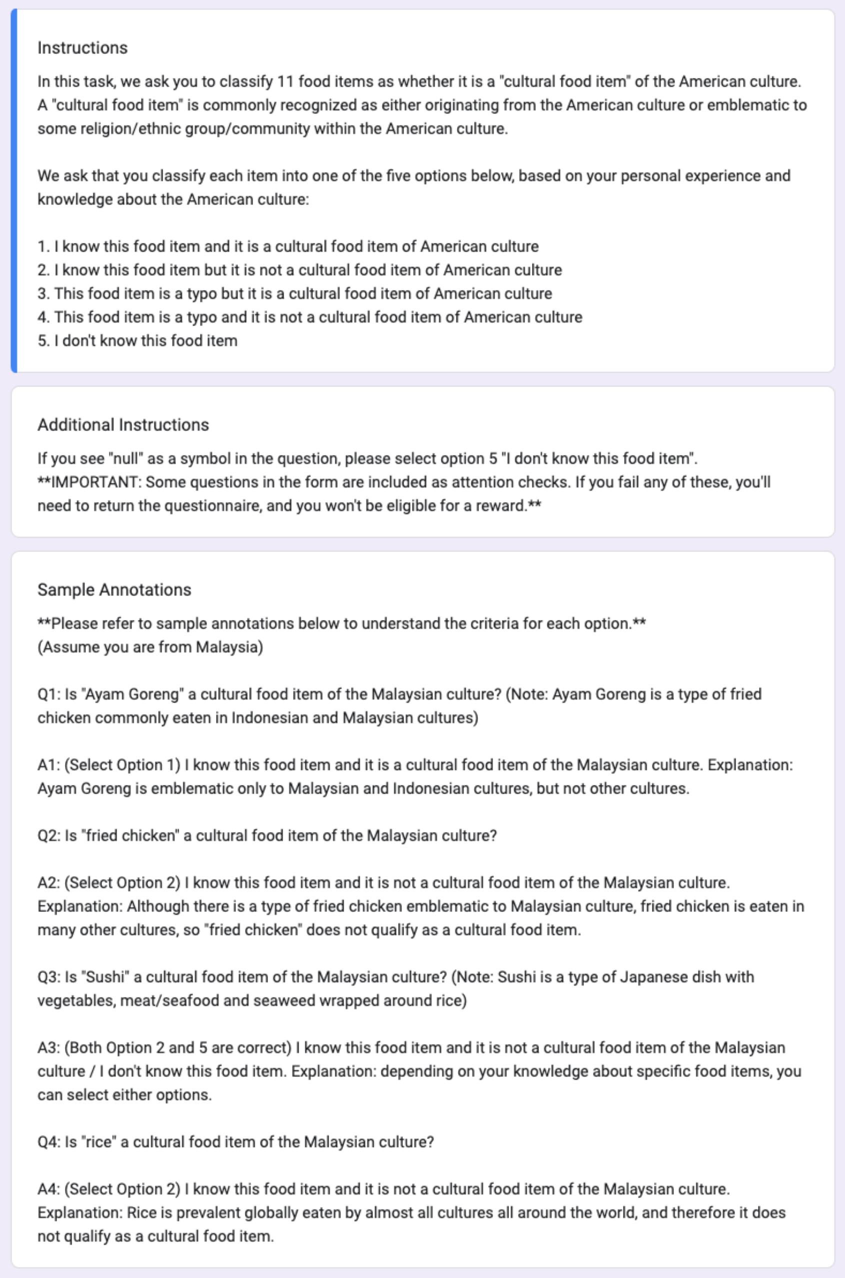
D Human Annotation Setup using Prolific

We designed a human annotation task using Google Forms, automatically populated via Google Apps Script with symbols related to food and clothing from eight different cultures. Figure 9 provides an overview of the form setup, while Figure 10 shows an example of a question where participants were asked to evaluate whether a specific food is a cultural food item of some culture. Annotators were required to select the most appropriate classification based on their knowledge of the culture in question. This process enabled us to collect reliable data regarding culturally emblematic food and clothing items.

E Ablation Study

E.1 Ablation on Hyperparameters

In the original design of our decoding process, multinomial sampling was employed with a set of specified hyperparameters: *temperature=1.0*, *top_p=0.95*, *top_k=50*, *max_tokens=30*, and *num_return_sequences=100*. The stopping criterion was established as the period ('.') character. To explore the impact of these parameters on the generation results, an ablation study was conducted where *top_k* values of 20 and 80, and *temperature* values of 0.75 and 1.25 were tested against the



Instructions

In this task, we ask you to classify 11 food items as whether it is a "cultural food item" of the American culture. A "cultural food item" is commonly recognized as either originating from the American culture or emblematic to some religion/ethnic group/community within the American culture.

We ask that you classify each item into one of the five options below, based on your personal experience and knowledge about the American culture:

1. I know this food item and it is a cultural food item of American culture
2. I know this food item but it is not a cultural food item of American culture
3. This food item is a typo but it is a cultural food item of American culture
4. This food item is a typo and it is not a cultural food item of American culture
5. I don't know this food item

Additional Instructions

If you see "null" as a symbol in the question, please select option 5 "I don't know this food item".
IMPORTANT: Some questions in the form are included as attention checks. If you fail any of these, you'll need to return the questionnaire, and you won't be eligible for a reward.

Sample Annotations

Please refer to sample annotations below to understand the criteria for each option.
(Assume you are from Malaysia)

Q1: Is "Ayam Goreng" a cultural food item of the Malaysian culture? (Note: Ayam Goreng is a type of fried chicken commonly eaten in Indonesian and Malaysian cultures)

A1: (Select Option 1) I know this food item and it is a cultural food item of the Malaysian culture. Explanation: Ayam Goreng is emblematic only to Malaysian and Indonesian cultures, but not other cultures.

Q2: Is "fried chicken" a cultural food item of the Malaysian culture?

A2: (Select Option 2) I know this food item and it is not a cultural food item of the Malaysian culture. Explanation: Although there is a type of fried chicken emblematic to Malaysian culture, fried chicken is eaten in many other cultures, so "fried chicken" does not qualify as a cultural food item.

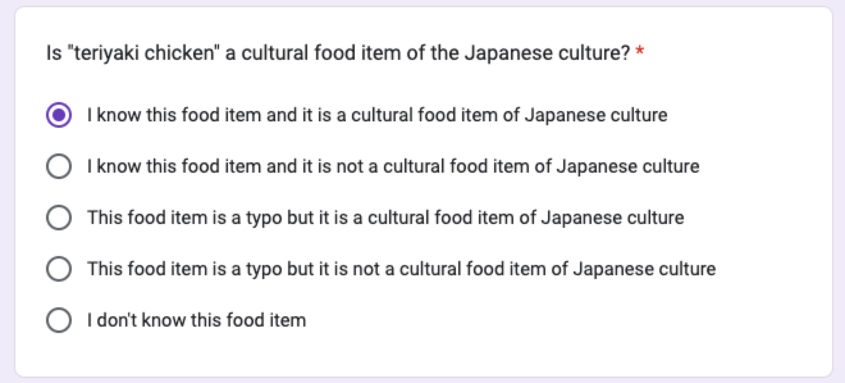
Q3: Is "Sushi" a cultural food item of the Malaysian culture? (Note: Sushi is a type of Japanese dish with vegetables, meat/seafood and seaweed wrapped around rice)

A3: (Both Option 2 and 5 are correct) I know this food item and it is not a cultural food item of the Malaysian culture / I don't know this food item. Explanation: depending on your knowledge about specific food items, you can select either options.

Q4: Is "rice" a cultural food item of the Malaysian culture?

A4: (Select Option 2) I know this food item and it is not a cultural food item of the Malaysian culture. Explanation: Rice is prevalent globally eaten by almost all cultures all around the world, and therefore it does not qualify as a cultural food item.

Figure 9: Example of Google Form Used for Cultural Food Annotation



Is "teriyaki chicken" a cultural food item of the Japanese culture? *

● I know this food item and it is a cultural food item of Japanese culture

○ I know this food item and it is not a cultural food item of Japanese culture

○ This food item is a typo but it is a cultural food item of Japanese culture

○ This food item is a typo but it is not a cultural food item of Japanese culture

○ I don't know this food item

Figure 10: Sample Question from Google Form on Cultural Food Classification

Ordering	Food	w/ Clothing
From Top 10 (↑)	Morocco - 107 Bangladesh - 99 Iceland - 99 Sweden - 96 Ethiopia - 90	Azerbaijan - 97 Bolivia - 96 Chile - 91 India - 76 Kenya - 74
From Bottom 10 (↓)	France - 42 Singapore - 42 Britain - 38 Indonesia - 36 Australia - 35	Germany - 30 United States - 28 China - 26 Portugal - 24 France - 21

Table 6: Cultures chosen for ablating on OLMo-7B-0424 and their corresponding number of unique symbols

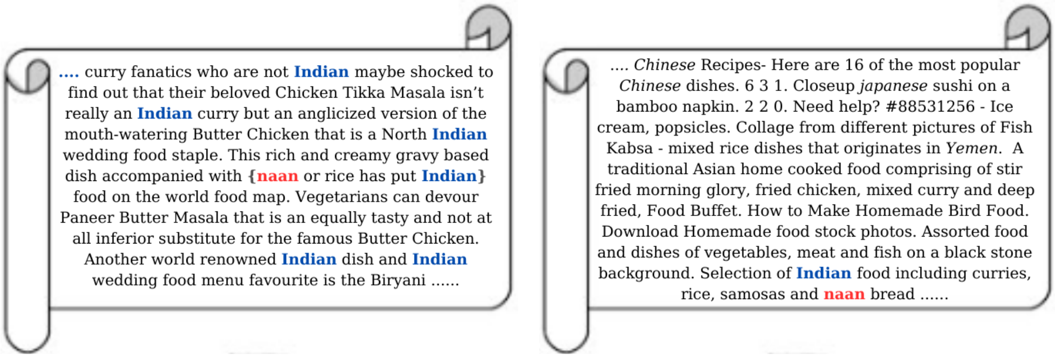
original settings. We observed an overlap coefficient of greater than 90% in all the four cases. This tells us that the sampling conditions did not cause or change our findings.

E.2 Ablation on OLMo-7B Variants

In order to verify that conclusions we find on OLMo-7B hold on other modalities, we reproduce some of the experiments on a newer variant of OLMo-7B, OLMo-7B-0424. We collect culture-conditioned generations for both food and clothing on OLMo-7B-0424, which is trained on Dolma 1.7. Although OLMo-7B-0424is the same model family as OLMo-7B, Dolma 1.7 contains pre-training documents that are not in Dolma 1.5, and OLMo-7B-0424 is trained with an updated algorithm from OLMo-7B. Other models supported by the WIMBD API, such as Pythia(Biderman et al., 2023), are not particularly capable of instruction following culture-conditioned generations, and therefore, analyzing their generations is less informative.

Due to the time constraints of the rebuttal, we only reproduce two main correlations in the main paper:

The number of cultures an independent symbol is generated for and the number of pretraining documents it appears in (Section 4.3) For OLMo-7B-0424, we obtain a moderate-to-strong correlation for both clothing (spearman $\rho = 0.507$, Kendall $\tau = 0.362$) and food (spearman $\rho = 0.416$, Kendall $\tau = 0.313$). Compared to OLMo-7B with clothing (spearman $\rho = 0.521$, Kendall $\tau = 0.367$) and food (spearman $\rho = 0.358$, Kendall $\tau = 0.260$), we see that even though the models and training data are different, the Spearman and Kendall correlations for food and clothing remain the same (both moderate-to-strong correlations). This means that the number of cultures an independent symbol was generated for and the number of pretraining documents it appears in is positively correlated, regardless of the model.



.... curry fanatics who are not **Indian** maybe shocked to find out that their beloved Chicken Tikka Masala isn't really an **Indian** curry but an anglicized version of the mouth-watering Butter Chicken that is a North **Indian** wedding food staple. This rich and creamy gravy based dish accompanied with {**naan** or rice has put **Indian**} food on the world food map. Vegetarians can devour Paneer Butter Masala that is an equally tasty and not at all inferior substitute for the famous Butter Chicken. Another world renowned **Indian** dish and **Indian** wedding food menu favourite is the Biryani ......

(a) d_{SNR}: 6.599 ; d_{TOK}: 4

.... *Chinese* Recipes- Here are 16 of the most popular *Chinese* dishes. 6 3 1. Closeup *japanese* sushi on a bamboo napkin. 2 2 0. Need help? #88531256 - Ice cream, popsicles. Collage from different pictures of Fish Kabsa - mixed rice dishes that originates in *Yemen*. A traditional Asian home cooked food comprising of stir fried morning glory, fried chicken, mixed curry and deep fried, Food Buffet. How to Make Homemade Bird Food. Download Homemade food stock photos. Assorted food and dishes of vegetables, meat and fish on a black stone background. Selection of **Indian** food including curries, rice, samosas and **naan** bread ......

(b) d_{SNR}: -0.982 ; d_{SENT}: 0

Figure 11: Examples of excerpts from relevant pre-training docs for Culture: "Indian" and Symbol: "Naan":

The number of memorized symbols for a culture and the number of pretraining documents it appears in (Section 4.1) For OLMo-7B-0424, exhaustively searching for all memorized symbols of all 110 cultures requires running MEMOED on all symbol-culture pairs, which is not feasible due to the rebuttal time constraint. Therefore, we select 10 cultures out of 110, 5 from the 10 cultures with the highest number of unique symbols generated by OLMo-7B-0424 and 5 from the 10 cultures with the lowest number of unique symbols generated by OLMo-7B-0424.

We obtain a moderate-to-strong correlation for both clothing (spearman $\rho = 0.591$, Kendall $\tau = 0.507$) and food (spearman $\rho = 0.829$, Kendall $\tau = 0.659$). Compared to OLMo-7B (on 110 cultures) with clothing (spearman $\rho = 0.540$, Kendall $\tau = 0.421$) and food (spearman $\rho = 0.670$, Kendall $\tau = 0.507$), we see that even though OLMo-7B-0424 is tested on smaller number of cultures, for both clothing and food, the correlation of OLMo-7B-0424. Therefore, the conclusion still holds that higher pretraining document counts of cultures increase the number of memorized symbols in culture-conditioned generations.

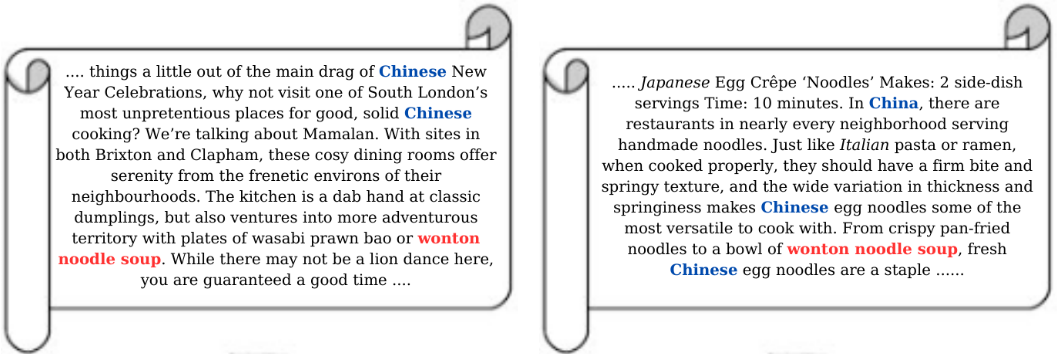
E.3 Ablation on Z-Score for MEMOED

We study whether selecting a different z-score threshold would change the conclusions of MEMOED on memorized symbols for all cultures. We perform an ablation study on setting the z-score to 2, which statistically means that the value is about 97.7 percentile. Empirically, a z-score below 2 does not indicate outliers, so we focus our ablation analysis only on cases where the z-score is 2.

When z=2, we get still get a moderate-to-strong correlation between 1) the number of memorized symbols for a culture and 2) the count of documents in which the culture appears in the pretraining corpora: for clothing, we obtain a spearman correlation of 0.569 and a Kendall correlation of 0.445; food food, we obtain a spearman correlation of 0.688 and a Kendall correlation of 0.519. This correlation is lower but similar to the original correlations found for z=2.6 (food: Spearman=0.670 and Kendall=0.507; clothing: Spearman=0.540 and Kendall=0.421), showing that our conclusion on the relationship between a culture's memorized symbols and the culture's frequency in pretraining data is robust to different z-score threshold.

In addition, we examine how lowering the z-score from 2.6 to 2 changes memorized symbols discovered for each culture. We compare each metrics's agreement with human evaluation on clothing: when z=2.6, the weighted F1 score is 0.845, and when z=2, the weighted F1 score is 0.840. We can see that z = 2 has a slightly lower agreement with human categorization, suggesting that additional symbols that are marked as memorized symbols when z=2 are non-emblematic symbols according to human culture experts.

F Training Document Excerpts



.... things a little out of the main drag of **Chinese** New Year Celebrations, why not visit one of South London's most unpretentious places for good, solid **Chinese** cooking? We're talking about Mamalan. With sites in both Brixton and Clapham, these cosy dining rooms offer serenity from the frenetic environs of their neighbourhoods. The kitchen is a dab hand at classic dumplings, but also ventures into more adventurous territory with plates of wasabi prawn bao or **wonton noodle soup**. While there may not be a lion dance here, you are guaranteed a good time ....

..... *Japanese* Egg Crêpe 'Noodles' Makes: 2 side-dish servings Time: 10 minutes. In **China**, there are restaurants in nearly every neighborhood serving handmade noodles. Just like *Italian* pasta or ramen, when cooked properly, they should have a firm bite and springy texture, and the wide variation in thickness and springiness makes **Chinese** egg noodles some of the most versatile to cook with. From crispy pan-fried noodles to a bowl of **wonton noodle soup**, fresh **Chinese** egg noodles are a staple ......

(a) d_{SNR}: 5.584 ; d_{TOK}: 17

(b) d_{SNR}: -0.841 ; d_{SENT}: 0

Figure 12: Examples of excerpts from relevant pre-training docs for Culture: "Chinese" and Symbol: "Wonton Noodle Soup":

In this section, we present excerpts from the pre-training documents classified as contributory to a culture-symbol association using MEMOED's d_{SNR}, d_{TOK} and d_{SENT} metrics.

In Figure 11, we present excerpts from two pre-training documents classified as contributory to the association between the culture: *Indian* and the symbol: *Naan*. We also report the relevant metric scores used to determine this. For Figure 11a, since the d_{SNR} is greater than zero, the d_{TOK} metric is used to ascertain the classification of this document. As visible in the excerpt, the culture "Indian" appears numerous times and in close proximity to the symbol "naan". Additionally, upon seeing the remaining part of the excerpt, we see that it is talking about Indian food items which indicates the relevancy of this document towards the association. On the other hand, for Figure 11b, since the d_{SNR} is between 0 and -1, we use the d_{SENT} metric as explained in Section 3.3. We can observe similarly that although the ratio is less than zero, the document is not noisy and the local context is about Indian food item.

Similarly, in Figure 12, we present excerpts from two pre-training documents classified as contributory to the association between the culture: *Chinese* and the symbol: *Wonton Noodle Soup*. We can observe that the training document with a positive d_{SNR} is not really talking about Chinese food items but rather talks about a prominent Chinese festival *i.e.*Chinese New Year and mentions the food delicacies being prepared then. Thus, through this it contributes to the association between the culture and symbol. On the other hand, for the document with negative d_{SNR}, we observe a relatively high concentration of cultural mentions in this excerpt and on a global level, the topic being discussed is restaurants in China when the food cultural symbol is mentioned. Hence we see how this document potentially contributes to the culture-symbol association.

G Additional Results

G.1 Culture Overmemorization

To further evaluate cultural overmemorization across all 110 cultures, we obtain: (1) the percentage of a culture's responses that contain another culture's memorized symbols; (2) the frequency of overmemorization for each culture, *i.e.* how often is a culture's memorized symbol generated for some other culture. Additionally, we calculate the correlation between each culture's metrics (1) and (2) with the frequency of topic-relevant occurrences of that culture in the pre-training corpora.

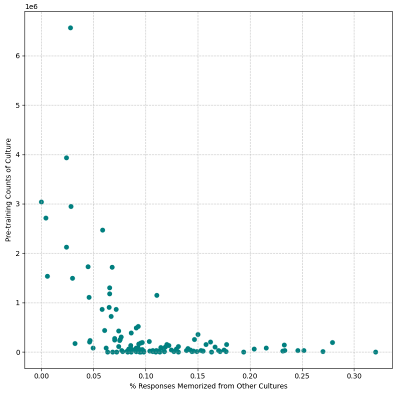


Figure 13: Correlation b/w Extent of Overmemorization and Pre-Training Counts for a Culture

Culture	**Overmemorizing Culture**	**Pre-Training Count Rank (/110)**
Trinbagonian	**American** (0.4%)	**101**
Macanese	**American** (0.5%)	**100**
Salvadoran	**American** (1%)	**99**
Zambian	**American** (0.6%)	**94**
Nicaraguan	**American** (0.4%)	85
Puertorriqueña	**American** (0.6%)	70
Egyptian	*Iranian* (2.9%)	27
Saudi	*Iranian* (6.2%)	45
Andorran	French (0.3%)	**110**
Hong Konger	French (0.6%)	38

Table 7: Cultures Identified from Leave-One-Out-Correlation

Topic	Keywords
favorite_music	music, song, songs, album, albums, band, bands, singer, singers, musician, musicians, genre, genres, concert, concerts
music_instrument	music instrument, music instruments, instrument, instruments
exercise_routine	exercise, routine, workout, sport, sports
favorite_show_or_movie	movie, movies, film, films, TV show, TV shows, TV series, cinema
food	food, foods, cuisine, cuisines, dish, dishes, meal, meals, recipe, recipes, menu, menus, breakfast, lunch, dinner, snack, snacks
picture	picture, pictures, painting, paintings, portrait, portraits
statue	statue, statues, sculpture, sculptures
clothing	clothing, clothes, apparel, garment, garments, outfit, outfits, attire, attires, dress, dresses, suit, suits, uniform, uniforms

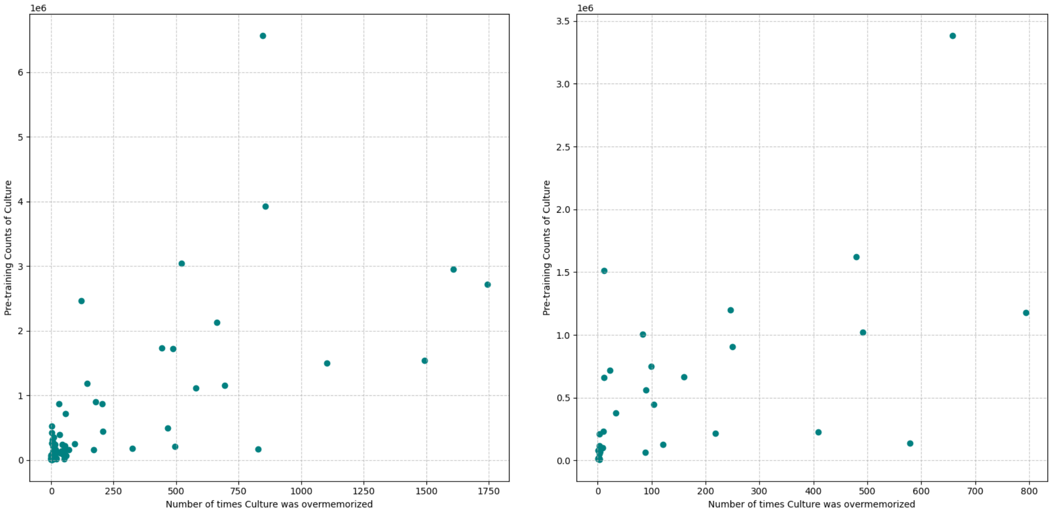
Table 8: Keyword list that we use to filter for topic-related pretraining documents.

For (1), we observe a moderate negative correlation for food (spearman $\rho = -0.521$, Kendall $\tau = -0.364$) indicating that cultures with high culture overmemorization tend to occur less-frequently in food-related pre-training documents. We have shown this correlation using a scatter plot in Figure 13. However for clothing, we observe a weak negative correlation (spearman $\rho = -0.099$, Kendall $\tau = -0.061$). To investigate this, we conducted a *leave-one-culture-out* experiment. In this analysis, we recalculated the correlations while systematically excluding one culture at a time. We then identified and listed the top ten cultures causing the highest variation. Notably, these cultures were predominantly those with significant overmemorization from regional cultures or those less frequently mentioned in the pre-training data, such as *Trinbagonian*. We have listed these ten cultures with the highest overmemorizing cultures in their responses (along with the percentage of responses being these overmemorized symbols) and their pre-training occurrence ranked out of all 110 cultures in Table 7. We observe that a majority of cultures have the highest cultural overmemorization from *America* while Egypt and Saudi have a significant percentage of their generations memorized from one culture *i.e.* Iran.

For (2), our observations indicate that 34 cultures related to clothing and 86 related to food were overmemorized at least once in the generations. Upon calculating correlations with these cultures, we observed moderate-to-high correlations for both clothing (Spearman $\rho = 0.763$, Kendall $\tau = 0.574$) and food (Spearman $\rho = 0.716$, Kendall $\tau = 0.531$). These results suggest that cultures frequently overmemorized are also those more commonly appearing in topic-related pre-training documents. We show this correlation through scatter plots for both clothing and food in Figure 14.

G.2 Results Overview

Continuing from Section 4.5, in this section we expand upon our findings and present some more results across the 110 cultures.



(a) Topic: Food

(b) Topic: Clothing

Figure 14: Cultural Overmemorization

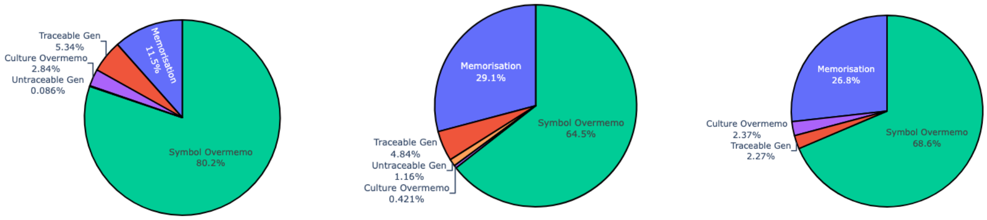
Ordering	w/ Memo	w/ Traceable Gen
Top 5 (↑)	Mexico India Japanese Morocco Nigeria	Trinidad Venezuela South Korea Morocco Georgia
Bottom 5 (↓)	Qatar South Africa Tajikistan Trinidad Yemen	Germany Japan United States Italy Denmark

Table 9: Memorization and Generalization Stats for Food

In Tables 9 and 10, we present the memorization and generalization statistics for food and clothing, respectively. Specifically, we provide the names of the top 5 and bottom 5 cultures, ranked by the percentage of their responses classified as either memorization or traceable generalization. Cultures with the highest percentage of memorized responses tend to correspond to those that appear more

Ordering	w/ Memo	w/ Traceable Gen
Top 5 (↑)	India Saudi Arabia Japan Pakistan Canada	Uruguay Venezuela Vietnam Yemen Zambia
Bottom 5 (↓)	Uruguay Venezuela Vietnam Yemen Zambia	Colombia Peru Nicargua Venezuela United States

Table 10: Memorization and Generalization Stats for Clothing



(a) China

(b) India

(c) Japan

Figure 15: Distributions of China, India and Japan responses for Food

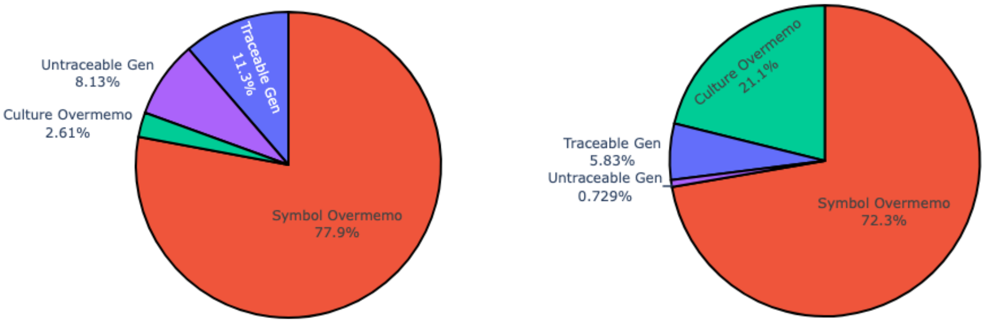
frequently in the pretraining dataset. However, notable exceptions exist, such as the culture *United States*, which, despite occurring frequently in the pretraining data and having a large number of memorized symbols, exhibits only 3.01% of its total responses as memorized, as shown in Figure 17a.

We also observe that a culture with a high percentage of memorized responses does not necessarily have a large number of unique memorized symbols. For instance, Pakistan ranks 4th in memorization count for the topic of clothing but has relatively few unique memorized symbols. This indicates that for some cultures, OLMo-7B tends to repeatedly generate the same memorized symbols when sampled multiple times. Additionally, Table 10 shows that the bottom 5 cultures, which have the lowest percentage of their responses classified as memorized, exhibit the highest percentage of traceable generalizations in their responses.

We further provide the distribution of additional cultures, similar to the analysis presented for Mexico and Trinidad in Section 4.5. Figure 15 illustrates the distribution of Chinese, Japanese, and Indian cultures for the topic of food. Notably, despite these three cultures being relatively high-frequency in the pretraining data, all exhibit very high symbol overmemorization rates, exceeding 60% in each case. Interestingly, we also observe considerable variance in the overall presence of memorization, ranging from almost 30% for India to only 11.5% for China. Additionally, all three cultures exhibit a relatively low percentage of culture overmemorization. This is likely due to their high frequency in the pretraining data, which results in their symbols being overmemorized in other, less frequently occurring cultures.

In Figure 16, we compare the distributions of two less-frequently occurring cultures, *i.e.*, Myanmar and Yemen, for the topic of clothing. We observe that, apart from exhibiting very high symbol overmemorization rates (greater than 70% in most cases), these cultures have no memorizations according to the classification provided by MEMOED. Consequently, there are no percentages of memorization recorded in their responses. Yemen, in particular, demonstrates a notably high percentage of cultural overmemorization, approximately 21.1%.

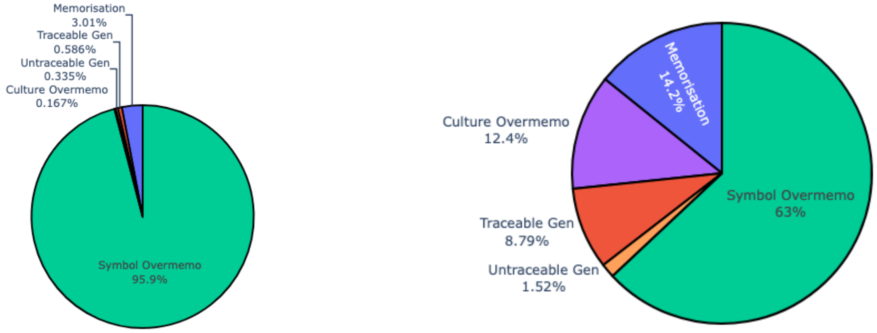
Finally, in Figure 17, we present the distributions for the USA and Saudi Arabia within the topic of clothing. The results for the USA are particularly striking, as it is one of the most frequently occurring cultures in the pretraining dataset, yet nearly 96% of its responses consist solely of symbol overmemorization. Despite containing a substantial number of unique memorized symbols, only 3% of its responses qualify as memorization. In contrast, Saudi Arabia exhibits greater diversity, with significant percentages of both memorization and cultural overmemorization in its generated outputs.



(a) Myanmmar

(b) Yemen

Figure 16: Clothing Stats - Mynammar and Yemen



(a) USA

(b) Saudi Arabia

Figure 17: Clothing Stats - USA and Saudi Arabia