Image Generated by Gemini

USC Viterbi School of Engineering
Thomas Lord Department of Computer Science

ink

Attributing Culture-Conditioned Generations to Pretraining Corpora

Huihan Li*, Arnav Goel*, Keyu He, Xiang Ren

ICLR 2025

Language models can learn to associate entities with cultures from pretraining data

For dinner, my Malaysian neighbor probably likes eating

- Nasi lemak
- Anchovies
- Vegetable salad
- Cucumbers



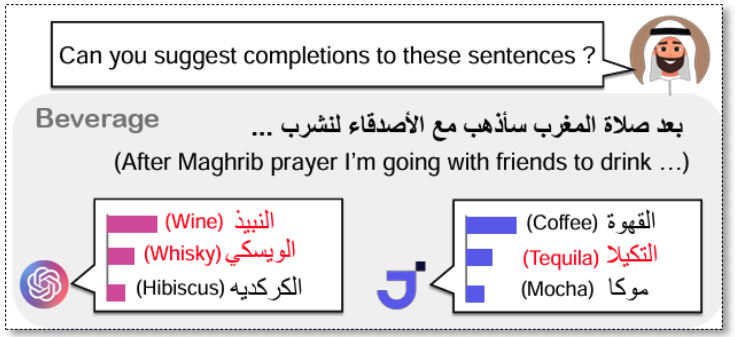

For dinner, my Indian neighbor probably likes eating

- Biryani
- Rice
- Dosa
- Mango lassi

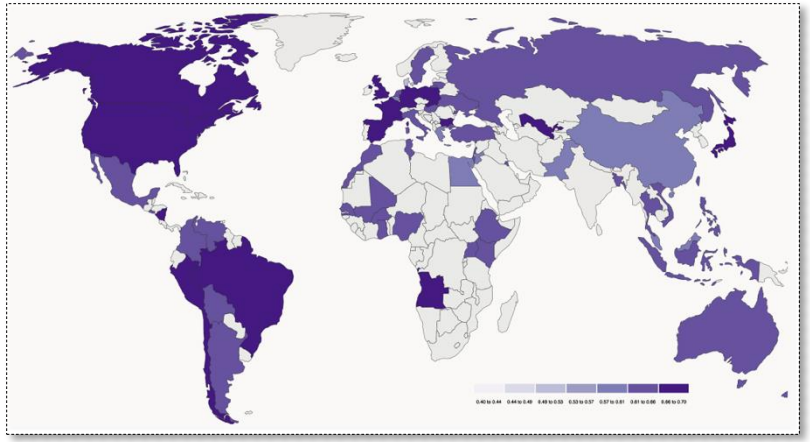



Language model acquires lots of culture-related knowledge during pretraining

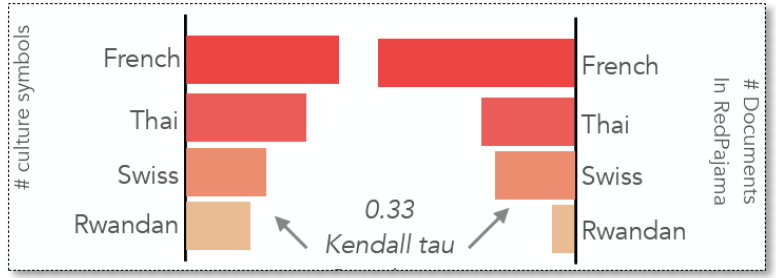
Previous works show that LLMs exhibit biases for less prevalent cultures



(Naous et al. 2023)
Model prefers Western-centric entities in Arabic context



(Anthropic. 2023)
LLM opinions are more similar to respondents from Western populations



(Li et al. 2024)
Culture with high pretraining data frequency has more diverse model generations

Our work (ICLR 2025)

- We investigate how pretraining leads to biased culture-conditioned generations by analyzing how models associate entities with cultures based on pretraining data patterns
- ***RQ1:***How can we *determine* if an entity is generated for a culture due to *memorization* of their association?
- ***RQ2:*** If not memorization, what *other factors* drive the model's association?
- ***RQ3:*** How are these types of associations tied to *pretraining data frequency imbalances*?

MEMOed: MEMOrization from pretraining document

If a culture co-occurs with the entity for a significant portion of the entity's pretraining documents, then the culture-entity's association is most likely memorized by the model.

Culture: India
Entity: dosa

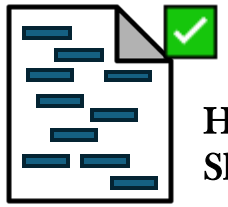
MEMOed

For brunch in **India**, the aforementioned **dosa** is a crispy rice lentil crepe that is traditionally served with

Kongu Nadu cuisine is predominantly South **Indian** with rice as its base … Idly, **dosa**, paniyaram and appam are popu

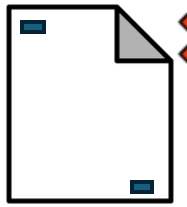
The **dosa** was great and the whole thing a real experience … D. said it reminded him of his trip to southern **India**.

1) Only retrieve pretraining document that are *relevant* to culture-entity association



contributes to memorization

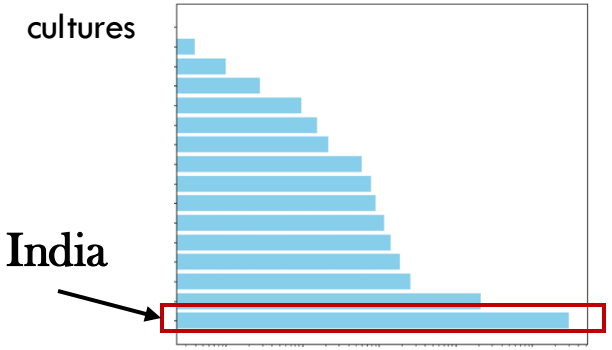
High n-gram density
Short culture-entity distance



no contribution to memorization

Low n-gram density
Long culture-entity distance

2) "Spike" in culture-entity co-occurrence distribution suggest that the culture-entity association is *much stronger than the rest*

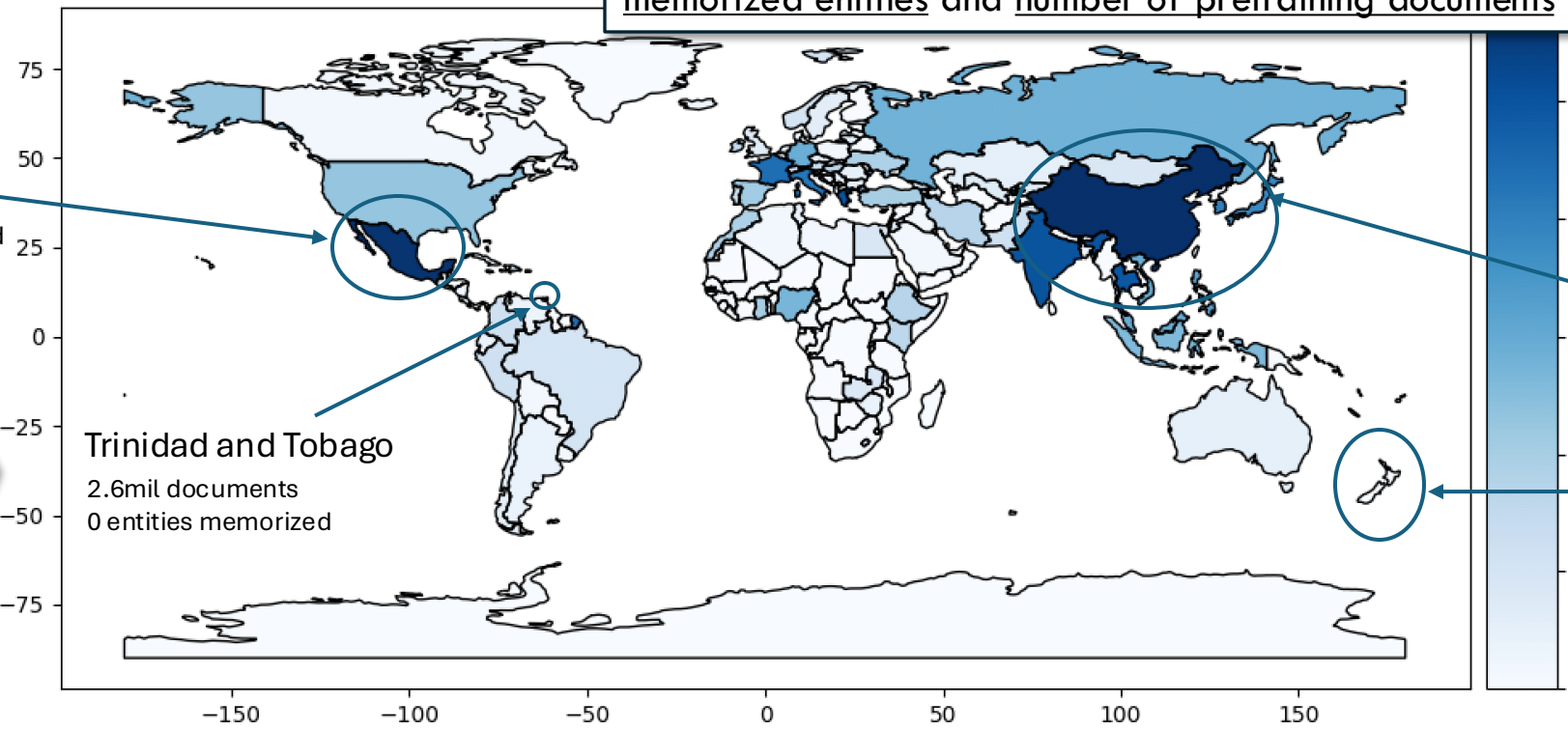


% culture-entity relevant documents / entity documents

"Dosa" is memorized for India because of distinguishably high co-occurrence

Number of memorized entity increases as the culture's pretraining frequency increases

We find a **strong positive** correlation ($\sigma = 0.670$) between number of memorized entities and number of pretraining documents for a culture



Mexico
60.3mil documents
57 entities memorized

Trinidad and Tobago
2.6mil documents
0 entities memorized

China
173.2mil documents
58 entities memorized

New Zealand
31.4mil documents
0 entities memorized

All food entities the model generates for Trinidad DO NOT associate to Trinidad from co-occurrence!

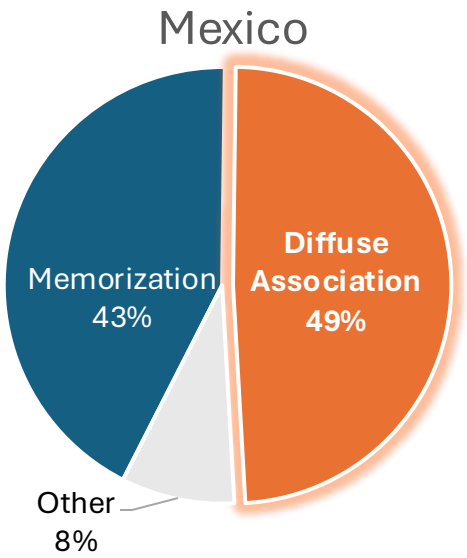
It is more difficult for models to memorize entities about underrepresented cultures

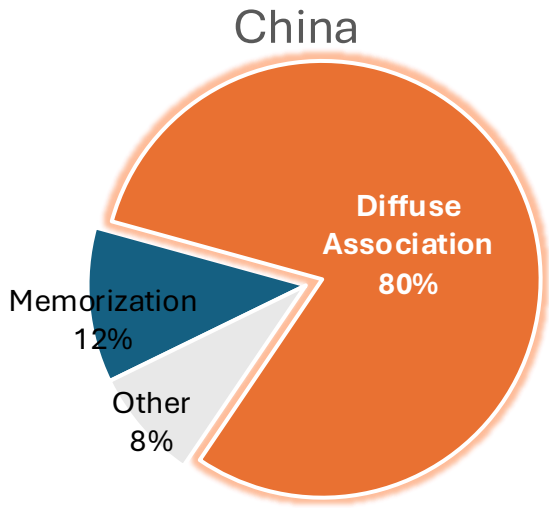
*We prompt OLMo-7B on 110 cultures, focusing on **food** and **clothing**, collecting 300 generations for each culture/topic.

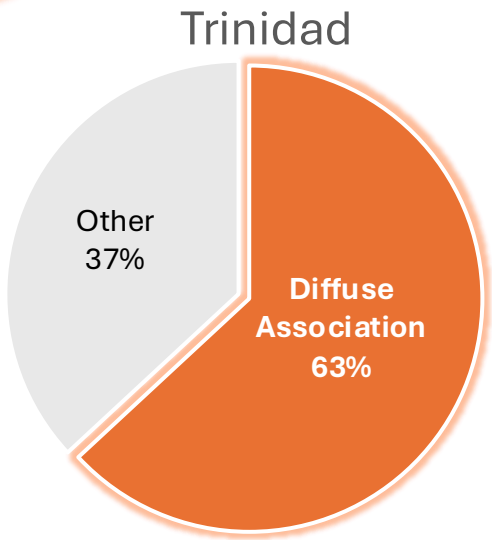
Diffuse Association entities: > 50% of generations for one culture

Examples: rice, t-shirt

- Compose of **4.1% (98)** total food entities
- But ALL are generated for **55+** cultures!
- **79%** of total generations are diffuse association entities

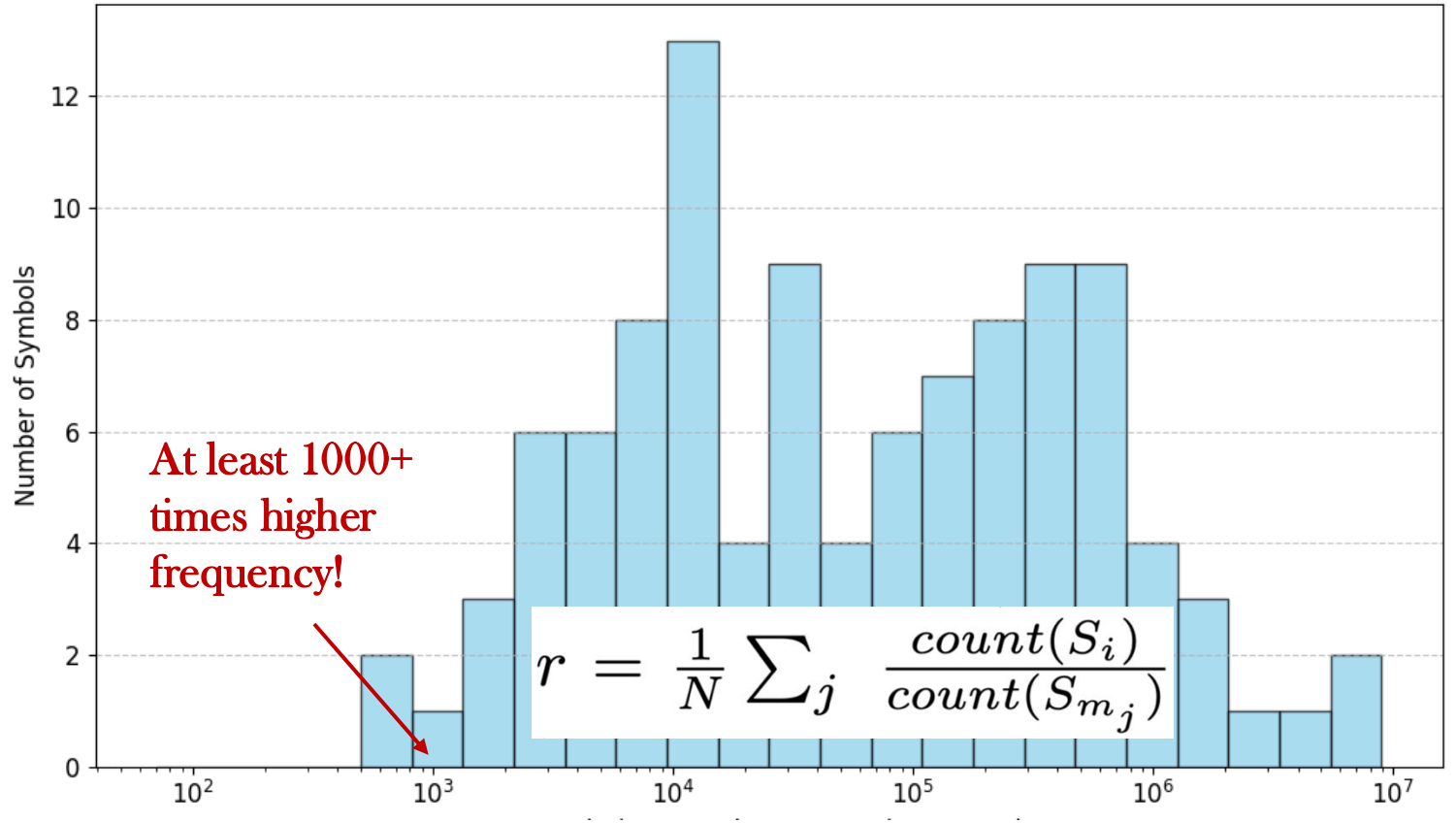






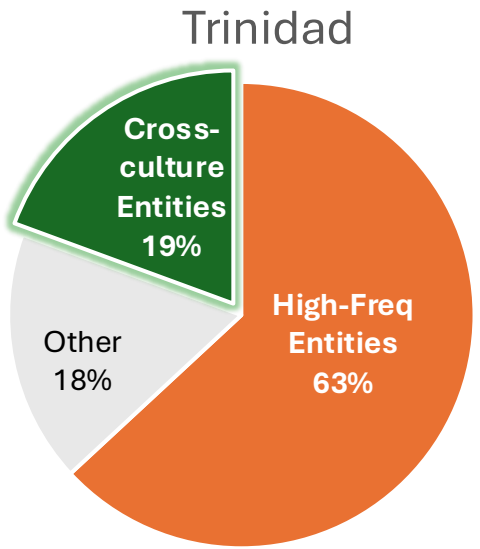
A small set of entities is prioritized over more culture-specific entities

Diffuse associations are high-frequency entities



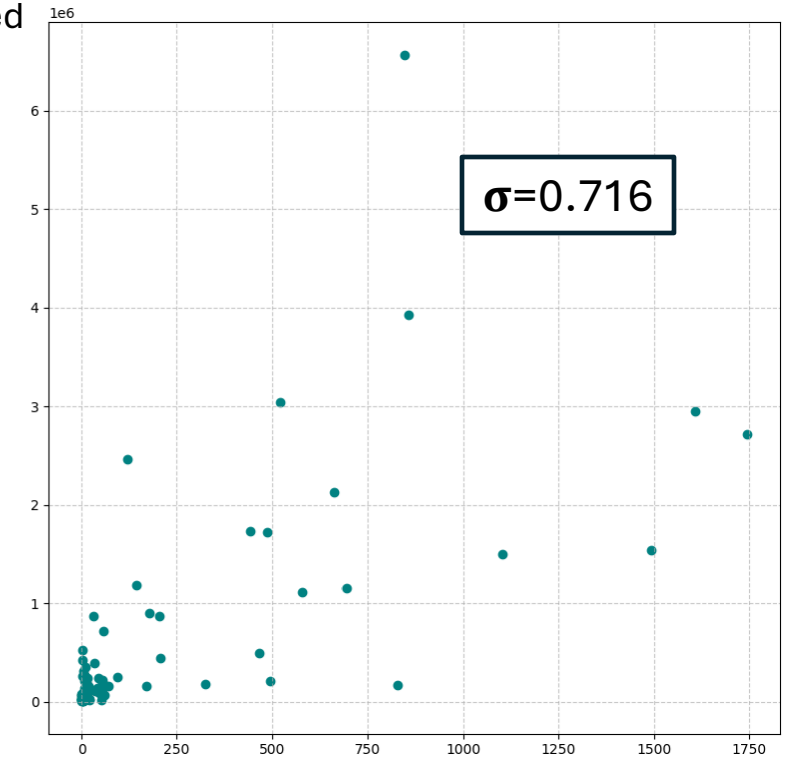
Model prefers to generate high-frequency entities independent of cultures over less frequent entities co-occurring with the culture

Entity memorized for one culture is generated for other cultures



- Chicken tikka masala (*India*) 
- Channa (*Pakistan*) 
- Boiled yam (*Nigeria*) 
- Chow mein (*China*) 

Food-related pretraining frequency



times a culture's memorized entity is generated for other cultures

High-frequency cultures' memorized entities are more likely sampled by models, even for other cultures

Key Takeaways

- We proposed the ***MEMOed framework*** to determine whether a generation for a culture arises from memorization.
- We find that language models are unable to ***reliably and evenly*** recall knowledge about global cultures in downstream generation, with ***high culture pretraining frequency*** positively influencing ***memorization***.
- We find that the model favors generating ***entities with extraordinarily high frequency*** or from ***high-frequency cultures***, regardless of the conditioned culture. This reflects biases toward frequent pretraining terms irrespective of relevance.

Thank you for listening!

- Code: https://github.com/huihanlhh/CultureGenAttr
- Paper: https://arxiv.org/abs/2412.20760