Global Gallery: The Fine Art of Painting Culture Portraits through Multilingual Instruction Tuning

Anjishnu Mukherjee[1], **Aylin Caliskan**[2], **Ziwei Zhu**[1], and **Antonios Anastasopoulos**[1]
[1]George Mason University, Fairfax, Virginia, USA
[2]University of Washington, Seattle, Washington, USA
{amukher6, zzhu20, antonis}@gmu.edu, aylin@uw.edu

Abstract

This study investigates the ability of Large Language Models (LLMs) to capture cultural nuances within various linguistic contexts. We focus on three areas: the effectiveness of language-specific instruction tuning, the influence of native language data pretraining, and methods to accurately extract cultural knowledge from LLMs. Our analysis uses the GeoMLaMA benchmark for multilingual commonsense knowledge and an adapted CAMeL dataset for cultural nuances, covering six languages and cultures. Results show that while targeted tuning and bilingual pretraining can improve cultural comprehension, they also reveal biases, especially in non-Western cultures. The findings emphasize the need for culturally diverse perspectives in LLM development for more inclusive language technology. [1]

1 Introduction

LLMs perform well at tasks centered around generating coherent text, benefiting from extensive pre-training on diverse datasets. However, they struggle with tasks requiring open-ended social reasoning, often producing biased responses (Parrish et al., 2022; Bender et al., 2021). Cultural influences play a crucial role in shaping social beliefs and behaviors, yet LLMs frequently reflect a Western bias due to unrepresentative training corpora (Weidinger et al., 2022). This bias can lead to cultural misrepresentation and insufficient understanding of non-Western cultural contexts in the models' outputs. Previous research has identified challenges in accurately depicting cultural nuances (Ramezani and Xu, 2023), perpetuating societal biases (Jakesch et al., 2023), and overlooking the subtleties of underrepresented cultures (Hutchinson et al., 2020), particularly in low-resource languages (Wibowo et al., 2023). To study cutural understanding in LLMs, we design the following research questions:

[1]Our code and data are available at this link

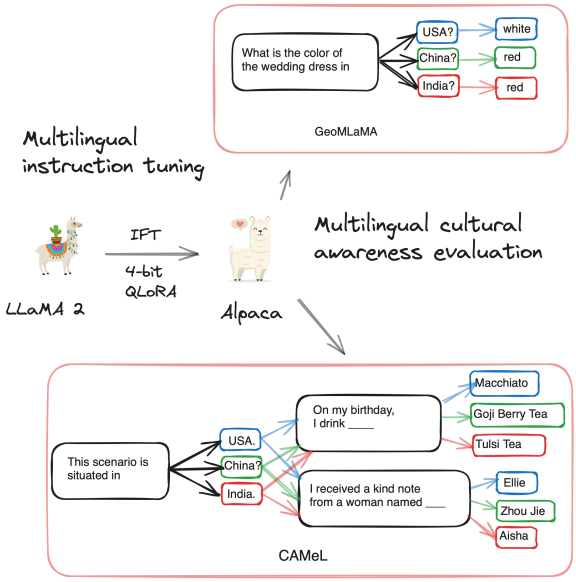

Figure 1: We instruction-tune LLaMA 2 in 5 non-English languages (Greek, Hindi, Persian, Swahili, Chinese) and evaluate both general cultural awareness as well as fine-grained multilingual cultural understanding.

RQ1 : Does instruction tuning on language-specific data enhance cultural knowledge?
RQ2a : Does pretraining on language-specific data enhance cultural knowledge?
RQ2b : What is the optimal approach for eliciting cultural information from LLMs?
RQ3 : Do LLMs understand the nuances of culture and what disparities exist across tangible cultural aspects?

2 Data

Our study looks at two datasets, GeoMLaMA and an adapted CAMeL dataset, to evaluate the cultural knowledge of LLMs at both broad and detailed levels. GeoMLaMA (Yin et al., 2022), extended to a QA format, spans six languages and cultures, including a new Greek variant, and features 900 questions covering 17 cultural topics. The adapted CAMeL dataset (Naous et al., 2023), originally for Arabic-Western cultural comparison, now encom-

passes the same six cultures and offers granular insight with prompts in nine categories for a nuanced analysis. To counteract position bias (Pezeshkpour and Hruschka, 2023), our multiple-choice QA scenarios further use randomized answer ordering.

3 Method

To assess the impact of language-specific instruction tuning on LLMs' cultural awareness, we utilized a 52k subset of English instructions from the Alpaca dataset (Taori et al., 2023), translated into five additional languages (Greek, Hindi, Persian, Swahili, and Chinesea) via the NLLB project's automatic translation system. This created the Alpaca-X dataset, with 'X' indicating the language version (e.g., Alpaca-en for English, Alpaca-hi for Hindi), all content-equivalent, only differing in terms of language. We then applied 4-bit QLoRA (Dettmers et al., 2023) for language-specific low-rank adapter training, integrating these adapters into a base LLM for each respective Alpaca-X variant. The hyperparameters, detailed in Appendix Table 12, ensures that each Alpaca-X model is fine-tuned to its cultural-linguistic context.

3.1 Experimental Settings

Our experimental framework is structured as follows to address the outlined research questions:

1. **RQ1:** We assess the impact of language-specific instruction tuning by comparing the performance of the English-tuned LLaMA 2 model against models tuned with Alpaca-X datasets.
2. **RQ2a:** The effectiveness of language-specific pretraining is evaluated using bilingual models with dedicated LoRA adapters for Chinese (Yi) and Swahili (Uliza).
3. **RQ2b:** We conduct an ablation study comparing the Swahili Alpaca adapter against a high-quality bilingual non-Alpaca adapter to examine the fine-tuning data's quality on cultural knowledge acquisition.

For a granular analysis of cultural aspects (**RQ3**), we utilize the CAMeL dataset across five experimental settings:

1. **Setting 1:** A multiple-choice question with options from various cultures to gauge understanding of specific cultural elements.
2. **Setting 2:** Questions with no correct cultural options to identify the model's cultural priors.
3. **Setting 3:** Multiple choices from the correct culture against a random incorrect option to test for the model's precision.
4. **Setting 4:** Questions with culturally correct but category-incorrect options to challenge the model's ability to discern between categories within the same culture.
5. **Setting 5:** Gendered questions to see if the model prioritizes cultural relevance over gender correctness.

4 Results

Our results (Appendix A) reveal that supervised finetuning (SFT) on language-specific data does not consistently enhance cultural knowledge across languages (**RQ1**), although LLMs outperform BERT and encoder-based models in English. The link between LLM parameter count and performance remains weak. Pretraining on language-specific data shows promise in improving cultural understanding (**RQ2a**), yet the significance of token counts used during pretraining versus finetuning requires further study. The most effective method for drawing out cultural knowledge (**RQ2b**) appears to be continued pretraining on high-quality data, coupled with targeted instruction tuning. When examining granular cultural elements (**RQ3**), LLMs demonstrate a skewed understanding influenced by pre-existing data distributions, and struggle with complex cultural discernment against distractors. Interestingly, LLMs tend to favor culturally relevant answers over grammatically gender-accurate ones when such conflicts arise.

5 Conclusion

This study on the cultural understanding of Large Language Models (LLMs) reveals significant variations in their ability to encapsulate diverse cultural nuances. Our investigations, leveraging the GeoMLaMA benchmark and the adapted CAMeL dataset, demonstrate that while language-specific instruction tuning and bilingual pretraining offer some improvements, they fall short of ensuring comprehensive cultural competence, particularly in non-Western contexts. The findings underscore the need for incorporating a wider range of cultural perspectives in LLM training and development, highlighting the importance of creating models that are not only linguistically adept but also culturally sensitive and globally inclusive.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. Copal-id: Indonesian language reasoning with local culture and nuances.

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.

A Appendix

We include a brief discussion about limitations of our approaches, along with the procedure followed for data collection, along with detailed results for each of the experiments for all the research questions, and the complete set of hyperparameters used in our experiments.

Limitations This study, is subject to certain limitations which are important to acknowledge:

1. The current methodology conceptualizes culture as a singular entity within a nation-state. This perspective, while useful for structured analysis, might not fully capture the rich diversity and complexity of modern societies, where multiple cultures and languages coexist within a single country. Future research could benefit from exploring more granular approaches that can effectively address this multifaceted nature of cultural identity.
2. The pretraining process lacks control over token distribution, contrasting with the controlled instructional data used in fine-tuning experiments. This could affect result interpretation. Future work should investigate the effects of smaller, high-quality datasets for controlled pre-training across languages.
3. Our experiments use 4-bit QLoRA for instruction tuning, and it's uncertain if results would differ with higher-bit configurations. Further research is needed to explore the impact of varying bit settings.
4. Evaluating large language models is an ongoing challenge within the field, and the methodology chosen for this study, while grounded in established research, has its strengths and limitations. This approach needs to be considered alongside alternative evaluation methods, each with their respective advantages and drawbacks, to suit specific use cases and research objectives.

Base Model	LoRA	Prompt
English	{lang} Alpaca	{lang}
English	English Alpaca	English
{lang}	{lang} Alpaca	{lang}
{lang}	Non-Alpaca LoRA	{lang}

Table 1: The four experimental combinations we test for RQ1 and RQ2. *lang* refers to language-specific variants of Alpaca or a language-specific prompt, translated from English.

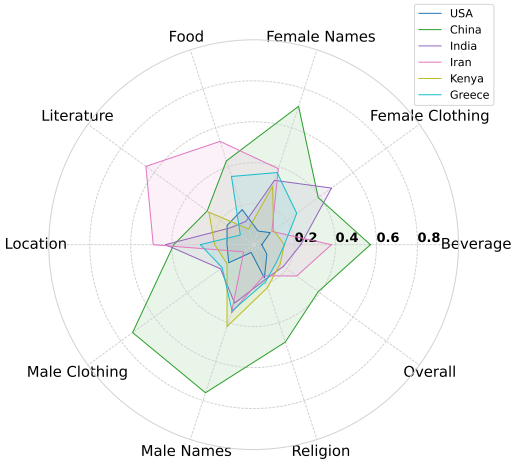

Figure 2: The 70B LLaMA 2 model shows strong performances for China and Iran across cultural concepts for different cultures.

Data collection from native speakers for adapted CAMeL dataset We provided native speakers with a list of words that we procured from different sources on the internet and from large language models as a base collection for each category that they are then asked to verify and correct with more appropriate targets for each category based on their lived experiences.
For the prompts, we follow a similar process, but this time we don't require country specific prompts, only category specific. The final set of prompts is decided by agreement between the authors.

SFT lang	China	India	Iran	Kenya	Greece
Results from GeoMLaMA benchmark					
(mBERT)	0.30	0.41	0.21	0.30	-
(XLMR-L)	0.37	0.37	0.37	0.32	-
Prompt language: english					
eng (7)	0.50	0.39	0.24	0.31	0.34
eng (13)	0.54	0.42	0.31	0.28	0.34
eng (70)	0.46	0.45	0.28	0.28	0.38
Prompt language: {lang}					
{lang} (7)	0.25	0.39	0.31	0.31	0.28
{lang} (13)	0.32	0.36	0.28	0.34	0.28
{lang} (70)	0.39	0.33	0.14	0.34	0.34

Table 2: Instruction tuning on language specific data does not consistently enhance cultural knowledge across languages and cultures. The numbers 7, 13 and 70 correspond to the model sizes in billions of parameters. The metric is the GeoMLaMA benchmark metric on a scale of 0-1 with higher being better.

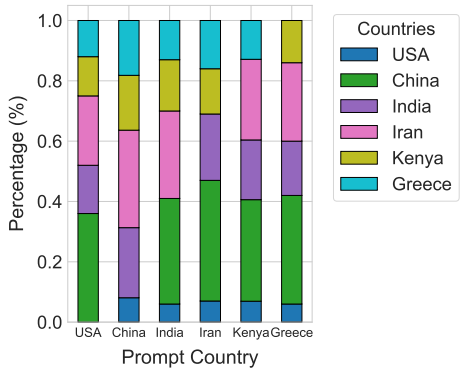

Figure 3: The distribution of countries chosen by the 70B LLaMA 2 model without the question explicitly mentioning the chosen country shows a large percentage favouring China and Iran.

We note that this process has inherent biases for the group of people who perform the tasks, which might implicitly show up in the data in unobserved ways. Also, because two of the categories are about names of people, this may include information about someone's real name, but that would only be so, because it is a common name in some part of their country.

All annotators are demographically located in the USA and are between 25-40 years old. Other than the Hindi annotator who is female, all others identify as male. Also, we note that all annotators are either authors or close friends of authors who did not require any form of compensation.

Model	Size	China	Kenya
Prompt language : English			
LLaMA 2 + eng Alpaca	7	0.50	0.31
	13	0.54	0.28
	70	0.46	0.28
Yi + eng Alpaca	6	0.43	-
	34	0.39	-
Uliza + eng Alpaca	7	-	0.25
Uliza + {swa, eng} LoRA	7	-	0.31
Prompt language : Chinese/Swahili			
LLaMA 2 + zh/swa Alpaca	7	0.25	0.31
	13	0.32	0.34
	70	0.39	0.34
Yi + zh Alpaca	6	0.39	-
	34	0.54	-
Uliza + swa Alpaca	7	-	0.31
Uliza + {swa, eng} LoRA	7	-	0.41

Table 3: Pretraining on language specific data helps to improve cultural awareness. Bilingual non-alpaca finetuning along with bilingual continually pretrained model gives the most culturally appropriate responses when prompted in the respective native language.

Category	USA	China	India	Iran	Kenya	Greece
Beverage	0.56	0.31	0.67	0.40	0.50	0.63
Female Clothing	0.60	0.69	0.58	0.81	0.79	0.69
Female Names	0.89	0.87	0.97	0.82	0.92	0.85
Food	0.32	0.40	0.76	0.32	0.69	0.28
Literature	0.21	0.33	0.45	0.20	0.34	0.65
Location	0.81	0.88	0.76	0.72	0.84	0.81
Male Clothing	0.58	0.54	0.85	0.86	0.75	0.74
Male Names	0.94	0.85	0.97	0.85	0.93	0.87
Religion	0.51	0.55	0.81	0.72	0.53	0.66
Overall	0.60	0.61	0.76	0.65	0.70	0.69

Table 4: We measure the percentage of times that LLaMA 2 70B prefers an option from an incorrect category when provided with a single choice from the correct category paired with 3 incorrect ones. Ideally, this should be close to 0 if the model has true understanding.

Category	USA	China	India	Iran	Kenya	Greece
Female Clothing	0.40	0.83	0.59	0.31	0.44	0.54
Male Clothing	0.51	0.66	0.77	0.42	0.48	0.68
Female Names	0.14	0.90	0.58	0.61	0.66	0.58
Male Names	0.23	0.93	0.57	0.65	0.68	0.65
Overall	0.32	0.89	0.63	0.51	0.56	0.61

Table 5: LLaMA 2 70B prefers being culturally correct than being gender correct across cultures.

Prompt	Model	Size	US	China	India	Iran	Kenya	Greece	Overall
English	LLaMA 2 + English Alpaca	7B	0.28	0.50	0.39	0.24	0.31	0.34	0.34
		13B	0.31	0.54	0.42	0.31	0.28	0.34	0.37
		70B	0.31	0.46	0.45	0.28	0.28	0.38	0.36
	Yi + English Alpaca	6B	0.52	0.43	0.33	0.48	0.50	0.34	0.43
		34B	0.62	0.39	0.42	0.45	0.50	0.44	0.47
	Uliza + English Alpaca	7B	0.21	0.50	0.39	0.17	0.25	0.31	0.31
	Uliza + {Swahili, English} LoRA	7B	0.45	0.39	0.39	0.34	0.31	0.25	0.36
Hindi	LLaMA 2 + Hindi Alpaca	7B	0.28	0.46	0.39	0.34	0.25	0.41	0.36
		13B	0.24	0.36	0.36	0.28	0.31	0.38	0.32
		70B	0.24	0.46	0.33	0.28	0.34	0.38	0.34
Chinese	LLaMA 2 + Chinese Alpaca	7B	0.34	0.25	0.39	0.41	0.41	0.34	0.36
		13B	0.38	0.32	0.39	0.48	0.47	0.38	0.40
		70B	0.38	0.39	0.42	0.48	0.53	0.34	0.43
	Yi + Chinese alpaca	6B	0.38	0.39	0.45	0.34	0.25	0.31	0.36
		34B	0.55	0.54	0.55	0.45	0.44	0.53	0.51
Swahili	LLaMA 2 + Swahili Alpaca	7B	0.34	0.32	0.39	0.17	0.31	0.34	0.31
		13B	0.34	0.29	0.39	0.24	0.34	0.34	0.33
		70B	0.31	0.36	0.39	0.21	0.34	0.38	0.33
	Uliza + Swahili Alpaca	7B	0.31	0.46	0.45	0.28	0.31	0.38	0.37
	Uliza + {Swahili, English} LoRA	7B	0.38	0.32	0.36	0.48	0.41	0.34	0.38
Persian	LLaMA 2 + Persian Alpaca	7B	0.31	0.25	0.27	0.31	0.38	0.38	0.32
		13B	0.31	0.25	0.33	0.28	0.34	0.34	0.31
		70B	0.28	0.36	0.33	0.14	0.25	0.38	0.29
Greek	LLaMA 2 + Greek Alpaca	7B	0.17	0.21	0.27	0.10	0.25	0.28	0.22
		13B	0.21	0.21	0.30	0.17	0.28	0.28	0.24
		70B	0.28	0.21	0.33	0.14	0.22	0.34	0.25

Table 6: RQ1, RQ2: Cultural performance scores of various models on the GeoMLaMA benchmark. Values are between 0 and 1, higher is better.

Category	Size	USA	China	India	Iran	Kenya	Greece
Beverage	7	0.03	0.44	0.16	0.34	0.10	0.09
	13	0.04	0.49	0.16	0.35	0.11	0.09
	70	0.04	0.57	0.23	0.38	0.15	0.15
Female Clothing	7	0.09	0.24	0.42	0.08	0.12	0.13
	13	0.11	0.30	0.46	0.09	0.15	0.14
	70	0.10	0.39	0.47	0.11	0.12	0.26
Female Names	7	0.05	0.50	0.21	0.28	0.19	0.18
	13	0.05	0.52	0.32	0.38	0.24	0.33
	70	0.07	0.71	0.33	0.39	0.30	0.37
Food	7	0.06	0.28	0.06	0.47	0.04	0.23
	13	0.07	0.33	0.08	0.49	0.06	0.21
	70	0.18	0.43	0.12	0.53	0.08	0.35
Literature	7	0.10	0.29	0.10	0.34	0.17	0.07
	13	0.12	0.27	0.12	0.39	0.19	0.06
	70	0.16	0.28	0.14	0.65	0.27	0.08
Location	7	0.09	0.28	0.27	0.39	0.13	0.17
	13	0.07	0.35	0.36	0.43	0.18	0.23
	70	0.13	0.39	0.43	0.49	0.19	0.26
Male Clothing	7	0.08	0.62	0.11	0.06	0.17	0.11
	13	0.09	0.68	0.18	0.10	0.19	0.12
	70	0.15	0.73	0.20	0.06	0.16	0.19
Male Names	7	0.02	0.53	0.22	0.26	0.30	0.20
	13	0.04	0.58	0.30	0.33	0.38	0.30
	70	0.04	0.76	0.30	0.35	0.42	0.34
Religion	7	0.16	0.41	0.09	0.11	0.24	0.11
	13	0.15	0.51	0.14	0.14	0.23	0.14
	70	0.17	0.50	0.18	0.16	0.22	0.19
Overall	7	0.08	0.39	0.18	0.26	0.16	0.14
	13	0.08	0.39	0.18	0.26	0.16	0.14
	70	0.08	0.39	0.18	0.26	0.16	0.14

Table 7: RQ3: Setting1 Results (Default MCQ setting, single correct country choice provided) from the CAMeL benchmark.

Prompt	Size	USA	China	India	Iran	Kenya	Greece
USA	7	0.0	0.36	0.17	0.22	0.14	0.12
	13	0.0	0.33	0.18	0.22	0.15	0.12
	70	0.0	0.36	0.16	0.23	0.13	0.12
China	7	0.1	0.0	0.24	0.27	0.22	0.17
	13	0.09	0.0	0.25	0.29	0.21	0.16
	70	0.08	0.0	0.23	0.32	0.18	0.18
India	7	0.07	0.37	0.0	0.24	0.18	0.14
	13	0.07	0.33	0.0	0.27	0.19	0.13
	70	0.06	0.35	0.0	0.29	0.17	0.13
Iran	7	0.09	0.40	0.21	0.0	0.16	0.14
	13	0.08	0.37	0.24	0.0	0.17	0.14
	70	0.07	0.40	0.22	0.0	0.15	0.16
Kenya	7	0.08	0.37	0.19	0.23	0.0	0.13
	13	0.08	0.34	0.21	0.25	0.0	0.13
	70	0.07	0.34	0.20	0.27	0.0	0.13
Greece	7	0.08	0.37	0.18	0.22	0.15	0.0
	13	0.07	0.33	0.20	0.23	0.16	0.0
	70	0.06	0.36	0.18	0.26	0.14	0.0

Table 8: RQ3: Setting2 Results (Distribution of Countries chosen when correct country is not provided) from the CAMeL benchmark

Category	Size	USA	China	India	Iran	Kenya	Greece
Beverage	7	0.48	0.05	0.22	0.09	0.34	0.37
	13	0.47	0.06	0.24	0.1	0.29	0.32
	70	0.5	0.05	0.19	0.08	0.25	0.27
Female Clothing	7	0.35	0.12	0.04	0.4	0.27	0.29
	13	0.37	0.12	0.04	0.35	0.25	0.28
	70	0.36	0.1	0.06	0.36	0.3	0.19
Female Names	7	0.48	0.07	0.19	0.16	0.21	0.2
	13	0.46	0.11	0.14	0.15	0.21	0.19
	70	0.44	0.04	0.2	0.17	0.19	0.17
Food	7	0.38	0.13	0.37	0.04	0.47	0.13
	13	0.37	0.12	0.34	0.04	0.45	0.17
	70	0.31	0.11	0.32	0.05	0.43	0.14
Literature	7	0.3	0.14	0.34	0.1	0.22	0.44
	13	0.28	0.14	0.29	0.07	0.22	0.45
	70	0.32	0.14	0.32	0.05	0.19	0.43
Location	7	0.42	0.13	0.14	0.11	0.28	0.26
	13	0.48	0.11	0.12	0.12	0.25	0.24
	70	0.45	0.13	0.13	0.08	0.29	0.24
Male Clothing	7	0.3	0.02	0.3	0.39	0.28	0.32
	13	0.32	0.02	0.24	0.31	0.25	0.25
	70	0.28	0.02	0.23	0.35	0.29	0.22
Male Names	7	0.65	0.06	0.19	0.21	0.16	0.18
	13	0.67	0.09	0.15	0.2	0.14	0.2
	70	0.67	0.02	0.17	0.2	0.17	0.17
Religion	7	0.27	0.08	0.32	0.36	0.2	0.28
	13	0.27	0.04	0.28	0.3	0.18	0.29
	70	0.27	0.06	0.28	0.28	0.17	0.23
Overall	7	0.4	0.09	0.23	0.21	0.27	0.28
	13	0.41	0.09	0.2	0.18	0.25	0.27
	70	0.4	0.08	0.21	0.18	0.25	0.23

Table 9: RQ3: Setting3 Results from the CAMeL benchmark (How many times did Llama choose the single incorrect option ignoring the other correct options. This number should ideally be 0 for everything.)

Category	Llama_Size	USA	China	India	Iran	Kenya	Greece
Overall	7	0.68	0.72	0.52	0.67	0.79	0.75
	13	0.73	0.71	0.72	0.72	0.78	0.62
	70	0.6	0.61	0.76	0.65	0.7	0.69
beverage	7	0.61	0.47	0.34	0.44	0.76	0.74
	13	0.66	0.43	0.53	0.64	0.68	0.58
	70	0.56	0.31	0.67	0.4	0.5	0.63
female_clothing	7	0.65	0.83	0.37	0.83	0.85	0.77
	13	0.68	0.81	0.62	0.83	0.78	0.57
	70	0.6	0.69	0.58	0.81	0.79	0.69
female_names	7	0.92	0.98	0.8	0.78	0.96	0.93
	13	0.98	0.99	0.93	0.94	0.97	0.77
	70	0.89	0.87	0.97	0.82	0.92	0.85
food	7	0.52	0.58	0.3	0.34	0.8	0.35
	13	0.47	0.46	0.63	0.33	0.8	0.22
	70	0.32	0.4	0.76	0.32	0.69	0.28
literature	7	0.26	0.37	0.23	0.52	0.41	0.6
	13	0.4	0.49	0.38	0.39	0.42	0.56
	70	0.21	0.33	0.45	0.2	0.34	0.65
location	7	0.8	0.94	0.6	0.66	0.94	0.93
	13	0.94	0.93	0.83	0.75	0.97	0.74
	70	0.81	0.88	0.76	0.72	0.84	0.81
male_clothing	7	0.74	0.65	0.59	0.81	0.8	0.81
	13	0.78	0.55	0.82	0.83	0.78	0.6
	70	0.58	0.54	0.85	0.86	0.75	0.74
male_names	7	0.95	0.94	0.78	0.85	0.93	0.91
	13	0.99	0.99	0.93	0.88	0.93	0.87
	70	0.94	0.85	0.97	0.85	0.93	0.87
religion	7	0.69	0.71	0.63	0.76	0.63	0.71
	13	0.63	0.65	0.78	0.78	0.67	0.62
	70	0.51	0.55	0.81	0.72	0.53	0.66

Table 10: RQ3: Setting4 Results from the CAMeL benchmark (How many times did Llama choose an option from the incorrect category) (it was given 3 incorrect categories, 1 correct category) - Ideally this should be 0 for everything if llama understands what category we are asking about.

Category	Size	USA	China	India	Iran	Kenya	Greece
Female Clothing	7	0.37	0.73	0.53	0.26	0.44	0.43
	13	0.38	0.85	0.59	0.34	0.49	0.41
	70	0.4	0.83	0.59	0.31	0.44	0.54
Female Names	7	0.12	0.85	0.53	0.52	0.65	0.46
	13	0.14	0.8	0.64	0.59	0.69	0.51
	70	0.14	0.9	0.58	0.61	0.66	0.58
Male Clothing	7	0.51	0.64	0.79	0.36	0.54	0.59
	13	0.48	0.68	0.8	0.45	0.52	0.56
	70	0.51	0.66	0.77	0.42	0.48	0.68
Male Names	7	0.21	0.85	0.59	0.55	0.6	0.56
	13	0.22	0.82	0.61	0.62	0.57	0.56
	70	0.23	0.93	0.57	0.65	0.68	0.65
Overall	7	0.3	0.82	0.61	0.44	0.55	0.51
	13	0.3	0.8	0.66	0.52	0.57	0.51
	70	0.32	0.89	0.63	0.51	0.56	0.61

Table 11: RQ3: Setting5 Results for the CAMeL benchmark(How many times did Llama choose correct culture but incorrect gender?) (2 options were from correct culture but opposite gender, and 2 options were from incorrect culture but correct gender)

Parameter	Value
Random Seed	42
Number of Epochs	1 (for 34B or 70B models), 3 (for 6B, 7B, 13B models)
Bits and Bytes Config	
Load	4 bit
Quantization Type	nf4
DataType	bfloat16
Lora Config	
Lora Alpha	16
Lora Dropout	0.1
R	64
Bias	none
Training Arguments	
Per Device Train Batch Size	6 (1 A100 80GB GPU)
Gradient Accumulation Steps	2
Learning Rate	3e-4
Max Gradient Norm	0.3
Warmup Ratio	0.03
Learning Rate Scheduler	constant
Optimizer	32bit paged AdamW
Max Sequence Length	2048

Table 12: Hyperparameters used for Instruction tuning of the LLaMA 2 models