CULTUREINSTRUCT: Curating Multi-Cultural Instructions at Scale

Viet-Thanh Pham[1], **Zhuang Li**[2], **Lizhen Qu**[1], **Gholamreza Haffari**[1*]
[1]Department of Data Science & AI, Monash University, Australia
[2]Royal Melbourne Institute of Technology, Australia
thanh.pham1@monash.edu, zhuang.li@rmit.edu.au, lizhen.qu@monash.edu, gholamreza.haffari@monash.edu

Abstract

Large language models, despite their remarkable success in recent years, still exhibit severe cultural bias. Therefore, in this paper, we introduce CULTUREINSTRUCT[1], a large-scale instruction-tuning dataset designed to reduce cultural bias in LLMs. CULTUREINSTRUCT is constructed with an automatic pipeline, utilizing public web sources and a specialized LLM to generate instruction. Our data comprises 430K instructions, ranging from classic NLP tasks to complex reasoning. CULTUREINSTRUCT also covers 11 most relevant topics to cultural knowledge, making it highly diverse. Our experiments show that fine-tuning LLMs with CULTUREINSTRUCT results in consistent improvements across three types of cultural benchmarks, including (i) general cultural knowledge, (ii) human opinions and values, and (iii) linguistic cultural bias. Our best model, QWEN2-INSTRUCT 72B + CULTUREINSTRUCT, outperforms GPT-4O Mini and GPT-4O with 18.47% and 13.07% average relative improvements on cultural benchmarks.

1 Introduction

Natural Language Processing (NLP) has seen drastic development in recent years, driven largely by advancements in Large Language Models (LLMs). These models have demonstrated remarkable capabilities in understanding and generating human language across a wide array of tasks, from question-answering (Arefeen et al., 2024) to complex reasoning (Huang and Chang, 2023). However, a critical challenge of LLMs remains unsolved: the integration of cultural knowledge into these models to ensure they operate unbiased across diverse linguistic and cultural contexts.

In addressing the problem of cultural biases, some works have attempted to actively mitigate these biases and improve the performance of LLMs on culturally diverse tasks. Researchers have explored various methods, such as prompt engineering with culturally specific context (Wang et al., 2024; AlKhamissi et al., 2024; Li et al., 2024c) to dynamically steer models toward generating culturally appropriate responses in real-time applications. However, prompt engineering is notorious for its unreliability across different downstream tasks. Several works tried to continue pre-train LLMs on non-English languages (Lin and Chen, 2023; Pipatanakul et al., 2023) but this approach is time-consuming and very costly, despite its effectiveness. Thus, it is better to fine-tune LLMs with instructions related to cultural knowledge. Recent works have proposed instruction-tuning datasets that are culture-specific, but they are either limited to a small set of instruction tasks (Shi et al., 2024) or only cover one aspect of cultural knowledge, such as cultural norms (Fung et al., 2023) and human values (Li et al., 2024a,b). These datasets, despite their high value in mitigating cultural biases, are only able to solve specific challenges within limited domains of cultural understanding. This narrow focus restricts the generalization potential of LLMs, making them less capable of handling a wide range of cultural contexts.

To bridge this gap, we present CULTUREINSTRUCT, a large-scale instruction dataset of cultural knowledge. CultureInstruct covers various topics related to cultural knowledge, including cultural norms, human values, history, cuisine, art, and more. To construct the dataset, we propose to **generate cultural instructions automatically from web documents**. Specifically, we implement a four-step pipeline to produce CultureInstruct (Figure 1). At the first step - **Document Filtering**, we select culture-relevant documents from Dolma (Soldaini et al., 2024) based on keyword filtering and apply a fastText model (Joulin et al., 2016) trained on an aggregated set of multiple public cultural

*Corresponding author.
[1]Dataset is available at https://github.com/thanhpv2102/CultureInstruct.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 9207–9228
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

datasets to further filter the documents. Secondly, at the **Instruction Generation** step, we utilize an open-source LLM - BONITO (Nayak et al., 2024) and tune it to generate instruction samples from the collected documents. This set of samples then goes through the third step - **Instruction Diversification**. In the final step, **Data Decontamination**, we remove samples that closely resemble any of the evaluation sets.

We validate the effectiveness of CultureInstruct by fine-tuning multiple series of LLMs, including LLAMA-3.1-INSTRUCT (Meta, 2024), MISTRAL-0.3-INSTRUCT (Jiang et al., 2023), QWEN2-INSTRUCT (Yang et al., 2024), OLMO (Groeneveld et al., 2024), and GEMMA2-INSTRUCT (Team et al., 2024). Fine-tuning with CULTUREINSTRUCT outperforms the base LLMs significantly on multiple types of cultural benchmarks, including: (i) benchmarks for general cultural knowledge - CULTURALBENCH (Chiu et al., 2024), CANDLE (Nguyen et al., 2023), and ETICOR (Dwivedi et al., 2023), (ii) human values - GLOBALOPINIONQA (Durmus et al., 2024); (iii) linguistic cultural bias - CAMEL (Naous et al., 2024). These benchmarks are particularly relevant because they assess models on diverse aspects of cultural understanding. By evaluating models on such culturally sensitive tasks, we ensure that the fine-tuned models demonstrate not only general competence but also cultural inclusivity and fairness in global contexts. In summary, we make the following contributions:

- We propose an automated data construction pipeline for generating high-quality cultural instruction data from raw documents.

- Utilizing the proposed pipeline, we build CULTUREINSTRUCT, a large-scale dataset of 430K multicultural instructions. CULTUREINSTRUCT covers a wide range of 11 domains within cultural knowledge and includes multiple types of instruction tasks.

- Our experiments reveal that fine-tuning LLMs with CULTUREINSTRUCT benefits models on cultural-related tasks (CANDLE and CULTURALBENCH), reducing cultural biases of models in cross-lingual (CAMEL) and cross-cultural settings (ETICOR). CULTUREINSTRUCT also makes LLMs more aligned with human values (GLOBALOPINIONQA). Our best model, QWEN2-INSTRUCT 72B + CULTUREINSTRUCT, outperforms GPT-4O Mini and GPT-4O with 18.47% and 13.07% average relative improvements on CANDLE, CULTURALBENCH, and ETICOR.

2 Data Construction

In this section, we describe the pipeline for constructing CULTUREINSTRUCT. As illustrated in Figure 1, the proposed pipeline consists of four processes: (1) selecting culture-relevant documents from DOLMA (Soldaini et al., 2024), (2) generating instruction data from the collected documents, (3) diversifying the instructions, and (4) decontaminate the data from the chosen cultural benchmarks.

2.1 Document Filtering

The pipeline begins by filtering the documents from the DOLMA corpus, which consists of two layers, keyword filtering and fastText retrieval. In the first layer, unlike previous works which only target one single aspect of cultural knowledge (eg. cultural norms (Fung et al., 2023; Ziems et al., 2023a)), we target a wider set of 11 cultural-related topics, including: general cultural knowledge, art, cuisine, cultural norms, festivals and national events, history, language, literature, music, religion, and social life. We first categorize the raw documents into these topics by applying a heuristic with multiple sets of keywords, corresponding to different topics (Appendix A.1). In the second filtering layer, we train a fastText model with seed data of 10 cultural-related datasets (Appendix A.2) and use the model to retrieve relevant documents. Specifically, we compute the embeddings of the documents and compute the similarity to samples in the seed data to select documents. After two filtering steps, the number of DOLMA documents was reduced from an initial 2,532M to 12M after keyword filtering, and further down to 214K for instruction generation.

2.2 Instruction Generation

Instruction data is generated using a tuned version of BONITO (Nayak et al., 2024). This model takes a document with a choice between 16 NLP tasks as input and generates the corresponding prompts & answers. We observe that the original BONITO model was trained with a large number of irrelevant documents to cultural knowledge (eg. scientific topics), hence it hallucinates when we provide cultural documents as input. Therefore, we construct a set of cultural instruction data to continue fine-tuning BONITO. To do this, we sample 200

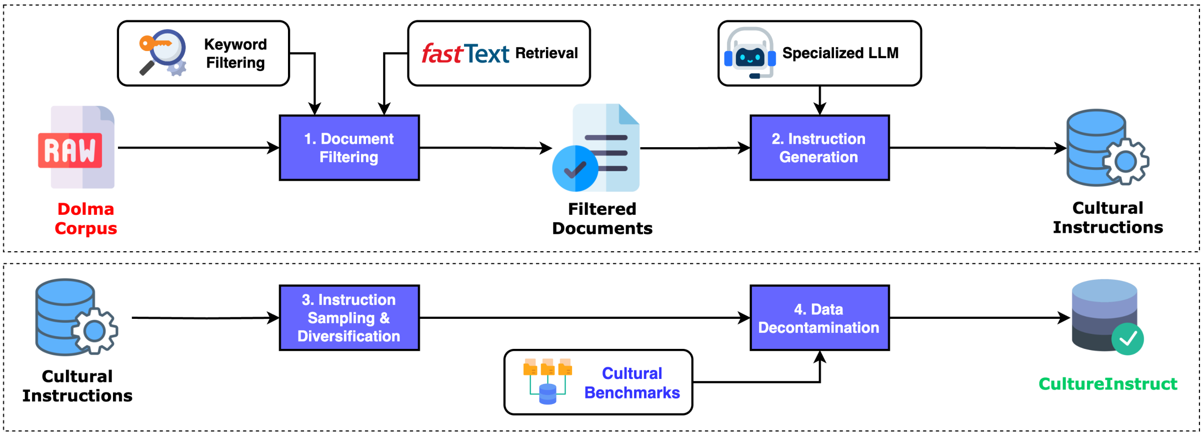

Figure 1: Illustration of our pipeline for building CULTUREINSTRUCT.

documents from each of the 11 topics, and prompt GPT-4o to generate the instruction data, resulting in 15,400 samples. The prompt template is given in Appendix B.1. We generate instructions following 6 NLP tasks (Extractive QA, Multiple-choice QA, Summarization, Textual Entailment, Topic Classification, and Sentence Completion), as well as an additional Chain-of-Thought reasoning task (Appendix B.3). The resulting data is used to continue fine-tuning BONITO. After fine-tuning, BONITO shows a large discrepancy in generation compared to its original version and can generate instructions effectively on cultural documents. Some comparisons of the models are described in Table 1, where the original BONITO hallucinates - revealing several problems, such as generating information not provided in the given context. Overall, 1.5M instructions are generated with our tuned BONITO.

2.3 Instruction Diversification

We found that many of the generated instructions have semantically similar or duplicated contents, due to the repeated information across web pages. To deduplicate and diversify the data, we first compute the embeddings[2] of the instruction data. We then calculate the cosine similarity scores between every pair of embeddings and consider a pair of embeddings is duplicated if their similarity score is high (> 0.90). From the filtered instructions, we further diversify them by applying the prototypical data pruning method (Sorscher et al., 2022). This approach starts by applying k-means clustering to the embedding space and progressively removing data points based on their distance to the closest cluster centroid. Following other works, we empirically set the value of k for k-means clustering to be the square root of the number of data points divided by two. We remove 10% of the data in each cluster. By doing this, the most semantically diverse samples remain. From the set of 1.5M raw instruction samples, more than 430K samples are kept after deduplication and diversification.

[2]We use `sentence-transformers/all-mpnet-base-v2` model to obtain the embeddings.

2.4 Data Decontamination

We perform data decontamination using two strategies, lexical-based and semantics-based:

- **Lexical-based** We follow Shao et al. (2024); Yue et al. (2024) to remove training samples that have a 10-gram matching with any of the samples in the benchmarks. For benchmark samples that are shorter than 10 grams but have at least 3 grams, exact matching is applied to filter out contaminated training samples.
- **Semantics-based** We compute the embeddings of samples using the same model in Section 2.3 with and remove samples that have cosine similarity of 0.9 or above with any of the test samples.

After the decontamination process, 99% of the data is retained, resulting in the final version of CULTUREINSTRUCT.

3 CULTUREINSTRUCT

3.1 Dataset Statistics

We present the statistics of CULTUREINSTRUCT and some other cultural training datasets in Table 2. Overall, CULTUREINSTRUCT covers a broader range of 11 domains, making it particularly valuable for applications that require a comprehensive

Document				
Vietnamese cuisine is extremely diverse, often divided into three main categories, each pertaining to Vietnam's three main regions (north, central and south). It uses very little oil and many vegetables, and is mainly based on rice and fish sauce. **Its characteristic flavors are sweet (sugar), spicy (Bird's eye chili), sour (lime), nuoc mam (fish sauce), and flavored by a variety of mint and basil.** Vietnam also has a large variety of noodles and noodle soups. Different regions invented typically different types of noodles, varying in shapes, tastes, colors, etc. One of the nation's most famous type of noodles is Pho, a type of noodle soup originating in **North Vietnam**, which consists of rice noodles and beef soup (sometimes chicken soup) with several other ingredients such as bean sprouts and scallions. Pho is meant to be savored, incorporating several different flavors: the sweet flavour of beef, sour lemons, salty fish sauce, and fresh vegetables.				
Generation Results				
Task	**Model**	**Generated Question**	**Generated Answer**	**Problems**
Extractive Q&A	BONITO	How long does it take to cook Pho?	A few minutes	Answer not mentioned in the document
	Tuned BONITO	What are the characteristic flavors of Vietnamese cuisine?	Its characteristic flavors are sweet (sugar), spicy (Bird's eye chili), sour (lime), nuoc mam (fish sauce), and flavored by a variety of mint and basil.	-
Summarization	BONITO	Guess a valid title for the document!	Vietnam	Title is too general
	Tuned BONITO	Give a title for the provided text	Vietnamese Cuisine	-
Sentence Completion	BONITO	Pho is traditionally served with ...	a plate of bean sprouts, basil, lemon wedges, and chili peppers on the side.	Answer not mentioned in the document
	Tuned BONITO	Pho is a type of noodle soup originating in ...	North Vietnam.	-

Table 1: Generation results of the tuned BONITO vs. original BONITO on 3 NLP tasks. **Blue** highlights the information successfully extracted by the tuned BONITO.

Dataset	**Domains Covered**	**Data Source**	**# of Samples**	**# of Countries**
NORMBANK (Ziems et al., 2023a)	Situational Social / Cultural Norms	Multiple databases to collect entities	155K	160
CANDLE (Nguyen et al., 2023)	Cultural Norms & Commonsense Knowledge	C4 Corpus (Dodge et al., 2021)	1.1M	176
CULTUREBANK (Shi et al., 2024)	Community-based Cultural Knowledge	Reddit, TikTok	23K	113
CULTUREINSTRUCT (Ours)	Cultural Norms, History, Cuisine, Art, Religion, Language, Music, Festivals, Literature, Social Life	DOLMA Corpus (Soldaini et al., 2024)	430K	183

Table 2: Comparisons of CULTUREINSTRUCT vs. other cultural datasets, namely NORMBANK, CANDLE, and CULTUREBANK.

cultural understanding. With 430K samples, CULTUREINSTRUCT provides a substantial amount of data as a robust resource for training models that require large and diverse datasets. While CANDLE is larger in terms of sample size (1.1M), its focus is only on cultural norms and commonsense knowledge, hence less diverse than CULTUREINSTRUCT. CULTUREINSTRUCT also covers data from 183 countries, making it highly diverse in cultural knowledge. Examples of CULTUREINSTRUCT are provided in Appendix B.2.

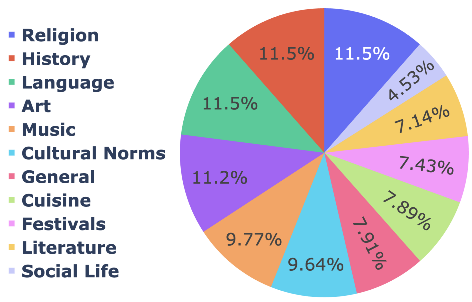

Figure 2: Topic distribution of CULTUREINSTRUCT.

3.2 Data Analysis

Topic Distribution. Figure 2 presents the topic distribution of CULTUREINSTRUCT, illustrating the proportions of various cultural topics represented within the dataset. Religion and History are the most prominent, followed by Cultural Norms and General topics. Notably, Social Life constitutes only 4.35% of the dataset, making it the least represented category. This is because data related to Social Life is more challenging to collect from public web corpora, as such content often resides in more culturally nuanced or user-generated contexts that are harder to filter. This data type needs human evaluation in the data processing pipeline to ensure quality, such as done in CULTUREBANK (Shi et al., 2024). Despite this limitation, the overall distribution reflects a well-balanced range of cultural topics, with most areas adequately represented.

Instruction Quality. To analyze the design choice of the instruction generation process in our pipeline, we compare the instruction quality of the original BONITO model and our tuned BONITO. We follow Xu et al. (2024) and prompt LLAMA-3.1-INSTRUCT 70B to judge the quality of the generated instructions (see Appendix B.4 for the prompt template and inference configuration). The quality distributions of the tuned BONITO and the original BONITO are illustrated in Figure 3. The majority of instructions generated by the tuned BONITO model are rated as "Excellent" (87.07%), while the ones from the original BONITO are mostly rated as "Average" (55.57%). To further validate the judgment made by LLAMA-3.1-INSTRUCT 70B, we conduct a manual check on a sub-sample of CULTUREINSTRUCT. LLAMA-3.1-INSTRUCT agrees with humans annotation at 70.93% accuracy rate; more details are described in Appendix C.1. We also analyze how instruction quality correlates with cultural knowledge relevance in Appendix C.2.

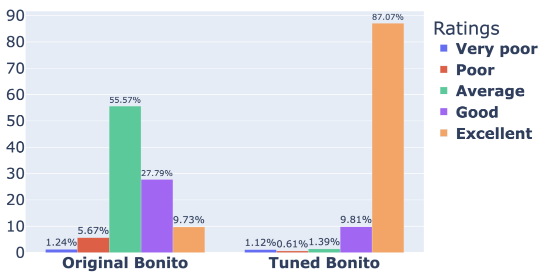

Figure 3: Instruction quality distribution of the original BONITO vs. the tuned BONITO.

Dataset Coverage. We analyze the coverage of CULTUREINSTRUCT vs other cultural training datasets, namely CANDLE, CULTUREBANK, and WORLD VALUE SURVEY. Specifically, we embed the samples in datasets with the embedding model in Section 2.3, then project the embedding to a two-dimensional space using t-SNE (van der Maaten and Hinton, 2008).

Figure 4 shows the t-SNE contour plot, where the Gaussian kernel is used for density estimation. For the raw scatter plot of t-SNE, see Appendix C.3. The representation of CULTUREINSTRUCT encompasses areas covered by other datasets. However, most of the area of CULTUREBANK is not covered by our dataset. This is because CULTUREBANK covers the Social Life topic, while this is the least

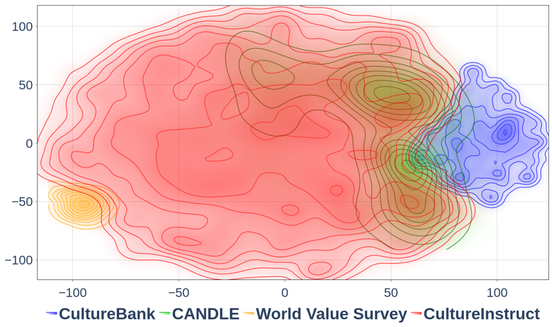

Figure 4: t-SNE contour plot of CULTUREINSTRUCT and other culture-related dataset, namely CANDLE, CULTUREBANK, WORLD VALUE SURVEY.

frequent topic in CULTUREINSTRUCT. Nevertheless, CULTUREINSTRUCT remains prominent as it includes instructions from diverse topics. We analyze this by computing the average Euclidean distance of each topic in CULTUREINSTRUCT against the other datasets. Results are presented in Table 3. The Literature topic has the largest distance to CULTUREBANK and CANDLE, indicating its under-representation in these datasets, while for WORLD VALUE SURVEY it is the Festivals topic. Other underrepresented topics include Music, History, and Religion.

Topic	CULTUREBANK	CANDLE	WORLD VALUE SURVEY
Religion	127.48	97.15	119.42
History	121.50	94.83	115.19
Cultural Norms	112.92	90.80	132.00
Cuisine	104.99	91.83	116.19
Art	114.19	89.37	119.71
Music	136.71	105.15	124.56
Festivals	103.85	79.80	**133.31**
Literature	**141.12**	**109.73**	128.98
Social Life	96.26	82.42	100.34
Language	118.58	90.54	125.70

Table 3: Average Euclidean Distance per topic in CULTUREINSTRUCT against CULTUREBANK, CANDLE, and WORLD VALUE SURVEY. **Bold** highlights the topics having the largest distances to each cultural dataset.

3.3 Additional Public Datasets

To enhance the quality of CULTUREINSTRUCT, we concatenate the constructed dataset with some other culture-related public datasets. Specifically, CULTUREBANK is the most important dataset to include, as it covers the Social Life topic that CULTUREINSTRUCT has very few samples of. Other culture-related datasets include: (i) a dataset of situational social norms - NORMBANK (Ziems et al., 2023a), (ii) a dataset of multicultural proverbs and sayings - MAPS (Liu et al., 2023), (iii) the training set of CANDLE (Nguyen et al., 2023), representing culture-related commonsense knowledge.

4 Experiments

In our experiments, we investigate to what extent CULTUREINSTRUCT improves the performance of various LLMs on cultural benchmarks, covering three aspects: (i) general cultural knowledge and cultural norms, (ii) linguistic cultural bias, and (iii) human values.

4.1 Models

Training Setups. The following models are used for training with CULTUREINSTRUCT: LLAMA-3.1-INSTRUCT (8B & 70B) (Meta, 2024), QWEN2-INSTRUCT (7B & 72B) (Yang et al., 2024), GEMMA2-INSTRUCT 9B (Team et al., 2024), MISTRAL-0.3-INSTRUCT 7B (Jiang et al., 2023), OLMO 7B (Groeneveld et al., 2024). For all models, we use QLoRA (Dettmers et al., 2023) during fine-tuning to save computational resources. Specifically, for models having less than 70B parameters, we use $r = 256$ and $\alpha = 256$, while $r = 128$ and $\alpha = 128$ are used for the larger models. All models are fine-tuned for 2 epochs on one A100-80B GPU.

Inference Setups. We evaluate all of the above-mentioned models and also include results from GPT-4O (version 2024-08-06) and GPT-4O Mini (version 2024-07-18). For these two models, we use the default decoding parameters for inference and in terms of open-sourced LLMs, we use greedy decoding for inference, except for the GLOBALOPINIONQA benchmark.

4.2 Cultural Benchmarks

We study the effect of fine-tuning on the following cultural benchmarks:

General Cultural Knowledge. We experiment with CANDLE (Nguyen et al., 2023) and CULTURALBENCH (Chiu et al., 2024) as multiple-choice question benchmarks. Both of these benchmarks cover multicultural knowledge, with CULTURALBENCH being a much harder benchmark due to its red-teaming approach in the data creation process. We also consider the ETICOR benchmark (Dwivedi et al., 2023) for evaluation. ETICOR covers region-specific etiquette for 5 regions: East Asia, India, Middle East & Africa, North America & Europe, and Latin America. With this data, the corresponding evaluation task is "Etiquette Sensitivity". Given a statement about etiquette, the task is to predict whether the statement is appropriate for a region.

Linguistic Cultural Bias. We evaluate LLMs on the CAMEL benchmark (Naous et al., 2024). Following their work, using CAMEL, we examine the performance in Arabic of models on the text-infilling task to show the model preference of Western vs. Arab entities. Specifically, we compute the probabilities of $[MASK]$ tokens in CAMEL prompts being Western or Arab entities. Then, the Cultural Bias Score (CBS) is used as the evaluation metric, where a lower score indicates less bias towards Western entities.

Human Values. GLOBALOPINIONQA (Durmus et al., 2024) is chosen as our benchmark for human values and opinions. This benchmark aggregates the Pew Global Attitudes Survey[3] and the World Values Survey[4], comprising around 2.5K questions. We follow their work to calculate the Jensen-Shannon distance between the human and model distributions, averaging over 5 prompts.

4.3 Main Results

General Cultural Knowledge. Table 4 shows the results of models on general cultural knowledge and norms benchmarks (CANDLE and CULTURALBENCH). Regarding the models ranging from 7B to 9B parameter sizes, fine-tuning with CULTUREINSTRUCT led to consistent improvements in both benchmarks. The models also showed marked improvements in the higher parameter category (70B–72B). Notably, QWEN2-INSTRUCT 72B, started with a relatively strong baseline performance (89.20%) but still exhibited improvement, reaching 90.80% accuracy. This model even outperforms GPT-4O and GPT-4O Mini on both benchmarks. This indicates that even models with high initial cultural competency can further benefit from fine-tuning with CULTUREINSTRUCT. Across both CANDLE and CULTURALBENCH benchmarks, LLMs with smaller parameter sizes (7B to 9B) tend to exhibit larger relative improvements after fine-tuning than the ones with larger parameter sizes (70B to 72B). This is due to the limitation in computational resources, we have to use a lower LoRA rank for larger LLMs, resulting in less forceful fine-tuning.

[3] https://www.pewresearch.org/feature/global-abortion/
[4] https://www.worldvaluessurvey.org/wvs.jsp

Model	Cultural Benchmarks		
	CANDLE (Accuracy ↑)	CULTURALBENCH (Accuracy ↑)	CAMEL (Cultural Bias Score ↓)
Parameter Size between 7B and 9B			
LLAMA-3.1-INSTRUCT 8B	72.20	48.81	51.42
+ CULTUREINSTRUCT	79.80	59.52	49.17
QWEN2-INSTRUCT 7B	77.80	47.62	47.51
+ CULTUREINSTRUCT	88.00	67.46	**41.78**
GEMMA2-INSTRUCT 9B	83.20	52.38	53.06
+ CULTUREINSTRUCT	83.40	67.44	47.99
MISTRAL-0.3-INSTRUCT 7B	76.60	44.84	50.38
+ CULTUREINSTRUCT	78.40	61.11	49.54
OLMO 7B	67.00	37.30	49.01
+ CULTUREINSTRUCT	70.20	46.83	42.02
Parameter Size between 70B and 72B			
LLAMA-3.1-INSTRUCT 70B	83.20	61.90	50.75
+ CULTUREINSTRUCT	89.80	71.43	47.60
QWEN2-INSTRUCT 72B	89.20	65.08	47.78
+ CULTUREINSTRUCT	**90.80**	**73.98**	42.20
Commercial LLM			
GPT-4O Mini (version 2024-07-18)	87.58	64.29	N/A
GPT-4O (version 2024-08-06)	89.38	73.00	N/A

Table 4: Performance of LLMs on cultural benchmarks. CANDLE and CULTURALBENCH belong to the **General Cultural Knowledge** benchmark. CAMEL represents the **Linguistic Cultural Bias** type. **Blue** highlights the best results in each column. Results of GPT-4O and GPT-4O Mini on CAMEL are unavailable due to classified token probabilities.

Regarding the ETICOR benchmark, fine-tuning with CULTUREINSTRUCT still shows consistent improvements across all of the regions Table 5, though with some exceptions in the Western region (America & Europe). For instance, LLAMA-3.1-INSTRUCT 8B initially achieves an F1 score of 85.35 for America & Europe, but after fine-tuning with CULTUREINSTRUCT, its performance slightly drops to 83.37. A similar pattern can be observed with QWEN2-INSTRUCT 7B, where its score decreases after fine-tuning. The results suggest that while fine-tuning with CULTUREINSTRUCT enhances cultural awareness of LLMs in non-Western regions, it may introduce slight trade-offs in regions where the model already performs well, like America & Europe. This could indicate that the training emphasizes broader multicultural knowledge, which benefits global regions but may occasionally conflict with region-specific nuances in Western contexts.

Linguistic Cultural Bias. The experimental results of the linguistic cultural bias benchmark - CAMEL are provided in Table 4. QWEN2-INSTRUCT $7B+$ CULTUREINSTRUCT achieves the best score of 41.78, indicating the lowest cultural bias. Similarly, OLMO $7B+$ CULTUREINSTRUCT shows a substantial reduction from 49.01 to 42.02 after fine-tuning. This demonstrates that the multi-cultural knowledge embedded in CULTUREINSTRUCT, despite being English-based, positively impacts the performance of LLMs in reducing bias on non-English tasks like CAMEL.

Human Values. The results from the GLOBALOPINIONQA benchmark in Table 6 show that fine-tuning with CULTUREINSTRUCT consistently improves model performance by reducing the Jensen-Shannon divergence, indicating better alignment with human values and opinions. Across different regions, models fine-tuned with CULTUREINSTRUCT exhibit lower divergence, which suggests that they are more aligned with the regional opinions assessed in the benchmark. These results reveal that CULTUREINSTRUCT helps models better understand and align with a variety of regional opinions, making them more culturally aware and responsive to human values across different countries.

Standard LLM Benchmarks. We report the results of models on several standard LLM benchmarks that are not related to cultural knowledge (HellaSwag (Zellers et al., 2019), CommonsenseQA (Talmor et al., 2019), and GPQA (Rein et al., 2024)) in Appendix D.1 to show the performance tradeoffs when fine-tuning with CULTUREINSTRUCT.

Model	EtiCor Benchmark (F1 ↑)				
	Latin America	India	Middle East	America & Europe	East Asia
Parameter Size between 7B and 9B					
Llama-3.1-Instruct 8B	60.90	67.73	59.53	85.35	63.70
+ CultureInstruct	71.35	69.98	72.98	83.37	72.73
Qwen2-Instruct 7B	68.35	76.75	72.54	84.84	72.16
+ CultureInstruct	77.43	**79.04**	78.70	**87.77**	77.79
Parameter Size between 70B and 72B					
Llama-3.1-Instruct 70B	62.97	73.70	61.67	85.93	65.52
+ CultureInstruct	73.20	77.98	76.03	85.52	76.81
Qwen2-Instruct 72B	69.18	77.85	76.05	86.57	76.05
+ CultureInstruct	**78.08**	78.77	**78.97**	85.37	**78.59**
Commercial LLM					
GPT-4o Mini (version 2024-07-18)	71.83	58.45	75.44	76.75	51.99
GPT-4o (version 2024-08-06)	74.81	59.80	78.16	79.96	54.33

Table 5: Performance of LLMs on the EtiCor benchmark (F1 scores). **Blue** highlights the best results in each column.

Model	GlobalOpinionQA Benchmark (Jensen–Shannon divergence ↓)			
	America	Japan	Germany	China
Parameter Size between 7B and 9B				
Llama-3.1-Instruct 8B	65.50	66.08	65.62	65.12
+ CultureInstruct	**55.71**	59.51	56.23	59.05
Qwen2-Instruct 7B	64.75	65.71	64.96	64.47
+ CultureInstruct	63.15	63.77	61.11	63.05
Parameter Size between 70B and 72B				
Llama-3.1-Instruct 70B	62.87	63.81	63.05	63.05
+ CultureInstruct	55.77	**56.29**	56.13	57.01
Qwen2-Instruct 72B	61.31	66.30	65.70	65.77
+ CultureInstruct	55.67	56.62	**56.01**	**56.96**

Table 6: Performance of LLMs on GlobalOpinionQA benchmark (Jensen–Shannon divergence). **Blue** highlights the best results from each column. Results of GPT-4o and GPT-4o Mini are unavailable due to classified token probabilities.

4.4 Effects of Fine-tuning on Model Outputs

In this section, we compare some base LLMs with their corresponding fine-tuned version with CultureInstruct to evaluate the hallucination and toxicity level after fine-tuning. To analyze the hallucination level of models, we evaluated the original and finetuned Llama-3.1-Instruct 8B and Qwen2-Instruct 7B on the TruthfulQA benchmark (Lin et al., 2022) - a dataset designed to specifically address the challenge of truthfulness and factual accuracy in AI-generated responses. Results on the subsets of the TruthfulQA benchmark are described in Table 7. After fine-tuning the models with CultureInstruct, the results demonstrate no significant performance decreases across any of the TruthfulQA subsets. Specifically, the accuracy on both MC1 and MC2 remains stable, and there is no noticeable decline in the ROUGE-L scores for generative outputs. These findings confirm that CultureInstruct does not negatively impact the factual accuracy or trustworthiness of the models, proving its compatibility with high-quality, reliable language generation.

Regarding the toxicity levels of models after fine-tuning, Table 8 shows the results of models on the ToxiGen benchmark (Hartvigsen et al., 2022). Similar to the TruthfulQA benchmark, after fine-tuning the models with CultureInstruct, the results reveal that there are no significant changes in performance, with accuracy remaining nearly identical. Specifically, Llama-3.1-Instruct 8B shows a minor increase, while Qwen2-Instruct 7B exhibits virtually no difference in accuracy. These results indicate that fine-tuning with CultureInstruct does not increase the toxicity level of the models, ensuring their outputs remain aligned with ethical and non-toxic standards.

5 Related Works

Cultural Knowledge Acquisition. In recent years, various cultural datasets have been constructed, and most of the works involve utilizing public sources and performing data synthe-

Model	TRUTHFULQA Benchmark		
	TruthfulQA MC1 (Acc ↑)	TruthfulQA MC2 (Acc ↑)	TruthfulQA Gen (ROUGE-L ↑)
LLAMA-3.1-INSTRUCT 8B	36.84	54.44	65.24
+ CULTUREINSTRUCT	36.79	54.31	65.18
QWEN2-INSTRUCT 7B	38.80	56.30	53.12
+ CULTUREINSTRUCT	38.77	56.14	52.98

Table 7: Performance of LLMs on the subsets of the TRUTHFULQA benchmark.

Model	TOXIGEN Benchmark (Acc ↑)
LLAMA-3.1-INSTRUCT 8B	45.85
+ CULTUREINSTRUCT	45.87
QWEN2-INSTRUCT 7B	47.77
+ CULTUREINSTRUCT	47.73

Table 8: Performance of models on the TOXIGEN Benchmark

sis for collecting cultural data. AlKhamissi et al. (2024); Masoud et al. (2023); Li et al. (2024a) utilize global opinion surveys, such as the World Value Survey and Hofstede survey to construct their cultural benchmarks. Similarly, CulturePark (Li et al., 2024b) proposed a multi-agent framework to synthesize data from the World Value Survey. Several studies have collected social norms from different cultures by prompting LLMs, including NormBank (Ziems et al., 2023b), NormSage (Fung et al., 2023), SocialDial (Zhan et al., 2023), ChineseNormBase (Qu et al.), and MulticulturalNormBase (Pham et al., 2024). CultureAtlas (Fung et al., 2024), CultureBank (Shi et al., 2024), and CANDLE (Nguyen et al., 2023) also sourced their data from public web pages with the help of robust NLP models for data filtering and synthesis. Our work closely follows the above-mentioned datasets, in which we also utilize public data sources and a specialized LLM for instruction generation. However, our work addresses the research gap of the previous works by including multiple facets of cultural knowledge, making CULTUREINSTRUCT more diverse.

Mitigating Cultural Bias. Several works have proposed approaches to mitigate the notorious cultural bias of LLMs, which can be categorized into three types: (i) prompt engineering (Wang et al., 2024; AlKhamissi et al., 2024; Li et al., 2024c; Zhan et al., 2024), (ii) continued pre-training (Lin and Chen, 2023; Pipatanakul et al., 2023), and (iii) instruction fine-tuning (Shi et al., 2024; Fung et al., 2023; Li et al., 2024a,b). In this work, we attempted to improve LLMs with the instruction fine-tuning approach, which is more consistent than prompt engineering and more cost-efficient than continued pre-training.

6 Conclusions

In this paper, we introduce CULTUREINSTRUCT, an instruction-tuning dataset designed specifically for improving cultural knowledge and reducing cultural bias of LLMs. CULTUREINSTRUCT consists of 430K meticulously-crafted instructions, covering 11 most relevant topics to cultural knowledge. Our experiments show that by fine-tuning with CULTUREINSTRUCT, various LLMs show consistent improvements on three types of cultural benchmarks, namely general cultural knowledge, linguistic cultural bias, and human values. In the future, we will extend the dataset to cover more instruction tasks that are specific to cultural knowledge, as well as extending CULTUREINSTRUCT to multilingual settings.

Acknowledgement

This work is partly supported by the ARC Future Fellowship FT190100039. This material is based on research sponsored by DARPA under agreement number HR001122C0029 (CCU Program). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

Limitations

The samples in CULTUREINSTRUCT are provided in English. Therefore, our dataset may not capture the cultural nuances appearing in language-specific characteristics. This is mainly due to the limitation of the BONITO model, as it can only generate English instructions, and extending it to multilingual may take a large amount of resources. In future work, we will attempt to apply translation models or extend the instruction generation model with multilingual capabilities.

Another limitation lies in the variety of models chosen for fine-tuning. Although these models are commonly used in other works, they still cannot fully represent the performance of other LLMs, such as those with less than 7B parameters or more than 72B parameters. Hence, we will expand the list of models to further test the usefulness of CULTUREINSTRUCT.

Ethical Statement

To regulate the use of CULTUREINSTRUCT and the models fine-tuned using it, we outline several ethical considerations and emphasize potential risks.

Misuse of Data. The primary goal of CULTUREINSTRUCT is to reduce cultural bias in LLMs by integrating diverse cultural knowledge. Despite our efforts to mitigate bias, this resource may contain some content that could be perceived as sensitive or controversial, particularly in the context of cultural differences. CULTUREINSTRUCT is constructed using publicly available web data and synthesized instructions generated by a specialized LLM. It is released for academic and research purposes and does not reflect the personal opinions or values of the authors. Any form of misuse, including employing this dataset to promote cultural discrimination or division, is strictly prohibited. Users are expected to adhere to the highest ethical standards, ensuring responsible use of this resource in alignment with research ethics. The authors and creators of CULTUREINSTRUCT hold no liability for misuse, misinterpretation, or unintended consequences of the dataset or models fine-tuned on it.

Risks in Data Generation. Since CULTUREINSTRUCT involves automatically generated instructions from a specialized LLM, there is a risk of inaccuracies or unintended bias within the dataset. We have tried to tune this LLM for better generation, but acknowledge that some instances of bias may remain. The creation of CULTUREINSTRUCT followed ethical guidelines to ensure that the dataset is inclusive and culturally diverse.

Potential Bias. While CULTUREINSTRUCT aims to minimize cultural bias in LLMs, it is impossible to eliminate bias entirely. The dataset and the models fine-tuned with it should be viewed as tools for research and improvement rather than final solutions to the issue of cultural bias in AI. We encourage further research, feedback, and iteration to continuously address and refine cultural fairness in AI systems.

By releasing CULTUREINSTRUCT, we aim to contribute to the responsible development of AI technologies that are more culturally aware. All users of this resource are expected to use it under ethical research practices, ensuring transparency and fairness.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Md Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. 2024. Leancontext: Cost-efficient domain-specific question answering using llms. *Natural Language Processing Journal*, 7:100065.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalteaming: Ai-assisted interactive red-teaming for challenging llms'(lack of) multicultural knowledge. *arXiv preprint arXiv:2404.06664*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *Preprint*, arXiv:2104.08758.

Esin Durmus, Karina Nguyen, 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. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.

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

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, et al. 2024. Olmo: Accelerating the science of language models. *Preprint*, arXiv:2402.00838.

Seungju Han, Junhyeok Kim, Jack Hessel, Liwei Jiang, Jiwan Chung, Yejin Son, Yejin Choi, and Youngjae Yu. 2023. Reading books is great, but not if you are driving! visually grounded reasoning about defeasible commonsense norms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 894–914, Singapore. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *Preprint*, arXiv:1612.03651.

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

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Huihan Li, Liwei Jiang, Jena D. Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024c. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *Preprint*, arXiv:2404.10199.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *Preprint*, arXiv:2311.17487.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *CoRR*, abs/2309.08591.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

Meta. 2024. Introducing llama 3.1: Our most capable models to date. `https://ai.meta.com/blog/meta-llama-3-1/`.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Nihal V. Nayak, Yiyang Nan, Avi Trost, and Stephen H. Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1907–1917, New York, NY, USA. Association for Computing Machinery.

Viet Pham, Shilin Qu, Farhad Moghimifar, Suraj Sharma, Yuan-Fang Li, Weiqing Wang, and Reza Haf. 2024. Multi-cultural norm base: Frame-based norm discovery in multi-cultural settings. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 24–35.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models. *Preprint*, arXiv:2312.13951.

Shilin Qu, Weiqing Wang, Xin Zhou, Haolan Zhan, Zhuang Li, Lizhen Qu, Linhao Luo, Yuan-Fang Li, and Gholamreza Haffari. Scalable frame-based construction of sociocultural normbases for socially-aware dialogues. *ACM Transactions on Multimedia Computing, Communications and Applications*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua Yu, Raya Horesh, Rogério Abreu de Paula, and Diyi yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *Preprint*, arXiv:2404.15238.

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. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *ArXiv*, abs/2406.08464.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, et al. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. 2024. Mammoth2: Scaling instructions from the web. *Preprint*, arXiv:2405.03548.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, et al. 2024. Renovi: A benchmark towards remediating norm violations in socio-cultural conversations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3104–3117.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, et al. 2023. Socialdial: A benchmark for socially-aware dialogue systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2712–2722.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023a. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023b. Normbank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776.

A Document Filtering

A.1 Keyword Filtering

Table 11 shows the list of keywords for each domain that we used in the Keyword Filtering step. Below is the heuristic for filtering documents:

- If the document contains less than n keywords from the "General Cultural Knowledge" or other domain lists, we label the document as "irrelevant".
- If the document contains less than n domain keywords and more than n general keywords, we label the document as "General culture knowledge".
- If the document contains more than n domain keywords, we label the document with the domain label with the highest number of keyword occurrences.
- Finally, we keep the document with the highest number of keyword occurrences.

In our approach, we empirically set $n = 3$ and apply the keyword filtering with the heuristic above.

A.2 FastText Retrieval

Training Configurations We utilize the fastText library to fine-tune a model with a vector dimension of 300, a learning rate of 0.1, a maximum n-gram length of 3, and a maximum number of word occurrences of 3. The checkpoint we used for fine-tuning is `crawl-300d-2M-subword` (Mikolov et al., 2018), which was pretrained on 600B tokens of the Common Crawl corpus[5].

Training Datasets To train the fastText model for retrieving relevant documents, we use the following 10 cultural-related datasets: SocialChem101 (Forbes et al., 2020), NormBank (Ziems et al., 2023a) CultureBank (Shi et al., 2024), CultureAtlas (Fung et al., 2024), MAPS (Liu et al., 2023), CALI (Culturally Aware Natural Language Inference) (Huang and Yang, 2023), CANDLE (Nguyen et al., 2023), NormLens (Han et al., 2023) , Social IQA (Sap et al., 2019), ETHICS (Hendrycks et al., 2021).

```
Given the following document:
{document}

Please generate some Q&A pairs that use
information from the document. Return them in a
JSONL format, without anything else. The Q&As
should be for the following NLP tasks:
- Summarization
- Extractive Question Answering
- Multiple-choice Question Answering
- Textual Entailment
- Topic Classification
- Sentence Completion

Please note that:
- The generated Q&A pairs should use the information
provided in the document.
- The JSONL should contain the following keys:
'task_name', 'question', 'answer'.
- For the "Multiple-choice Question Answering" task,
there should be another key named 'options' that
includes 4 answer choices (A, B, C, D).
- If the documents are not suitable for any of the tasks,
simply answer with a JSONL containing 'Not suitable' as
a value for a key named 'reason'.
```

Figure 5: Prompt template for generation instructions using GPT-4o.

B Instruction Generation

B.1 GPT-4o Prompt for Instruction Generation

Figure 5 shows the prompt used for generating instruction data using GPT-4o. The generated instructions are used to continue fine-tuning the BONITO model.

B.2 Generated Examples

We illustrate several examples of CULTUREINSTRUCT in Table 12. There are examples provided for 6 NLP tasks, namely Summarization, Extractive Question Answering, Multiple-choice Question Answering, Textual Entailment, Topic Classification, and Sentence Completion. The examples for the additional reasoning task are provided in Appendix B.3.

B.3 Chain-of-Thought Reasoning Task Generation

Figure 6 describes the prompt for generating Chain-of-Thought Reasoning instructions using GPT-4o. Table 13 shows some example instructions generated by our tuned BONITO model.

B.4 Instruction Quality Assessment

To assess the quality of the instructions generated by the tuned BONITO, we prompt LLAMA-3.1-

[5]https://commoncrawl.org/

Given the following document:
{document}

Please generate some Q&A pairs that use information from the document. Return them in a JSONL format, without anything else. Please note that:
- The generated Q&A pairs should use the information provided in the document.
- The JSONL should contain the following keys: 'task_name', 'question', 'answer'.
- The answers should contain a short introduction, followed by multi-step reasoning (provide a step-by-step explanation) before coming up with the conclusions - final answers.
- If the documents are not suitable for reasoning, simply answer with a JSONL containing 'Not suitable' as a value for a key named 'reason'.

Figure 6: Prompt template for generation Chain-of-Thought Reasoning instructions using GPT-4O.

INSTRUCT 70B with the prompt template shown in Figure 9. For each instruction generated by BONITO, we prompt LLAMA-3.1-INSTRUCT 70B 5 times and take the majority voting as the quality label.

C Additional Data Analysis

C.1 Assessing LLM Judgement on Instruction Quality

To assess the reliability of using LLAMA-3.1-INSTRUCT 70B for judging instruction quality (Section 2.2 and Section 3.2), we perform a manual evaluation of the judgments of the model. Firstly, we hired 5 graduate students for this task, so the annotation quality is better than standard crowd-sourcing workers. The pay rate for each annotator is 20 USD per hour. We then perform stratified sampling with the generated instructions based on the task type, and 234 samples are selected for labeling. We then prompt LLAMA-3.1-INSTRUCT 70B for its quality labels (instruction prompt template shown in Figure 9) and have each human annotator label the samples, following the identical instruction given to the model.

After completing the annotation process, we first collect the majority human selection for each sample. As the number of human annotators is odd, it is unnecessary to check for cases where there may be an equal number of selections - no majority selection. Then the model's reliability can be assessed by computing the accuracy of the model predictions w.r.t. human majority selections. **Within 234 samples, LLAMA-3.1-INSTRUCT agrees with humans on 166 samples, reaching 70.93% accuracy**. We further perform qualitative analysis on cases where there is disagreement between the model and human annotators.

Qualitative Analysis Table 14 shows some examples for analyzing human judgment and LLM judgment. In the first example, the answer contains information that is not provided in the context (highlighted in Red), hence the quality score should be "Poor", as labeled by human annotators. Example 3 follows the same pattern, where the LLM judge cannot detect the hallucinated content in the instructions. In the second example, where the task is to summarize the main subject of the provided text, the generated answer is missing several sections in the text (Geography, Transportation, Economy, Education, Sport), as well as containing some hallucinated sections (History, Culture). Human annotators rated this example as "Poor", while the LLM Judge could not detect the problem. However, in the last two examples, the judgments of LLM are better than human evaluators. Based on the explanation made by the model, it is clear that in this case, the LLM Judge successfully detected the problems in the generated instructions that human annotators cannot. This analysis highlights the strengths and weaknesses of LLM judgment, demonstrating that while it can outperform human evaluators in some cases, it still struggles to consistently identify hallucinations and missing content.

C.2 Assessing Instruction Quality & Culture Relevance

Figure 7: Instruction Quality and Cultural-Relevance distributions.

In this section, we analyze to what extend the instruction quality correlates with cultural knowledge relevance. Figure 7 presents a heatmap illustrating the joint distribution of Instruction Quality and

Cultural-Relevance ratings, expressed as percentages. To judge the instruction quality and cultural relevance, we prompt LLAMA-3.1-INSTRUCT 70B to obtain the quality labels (Figure 9) and the culture relevance labels (Figure 10. From the figure, a substantial portion of CULTUREINSTRUCT is concentrated in the higher rating categories ("good" and "excellent"). This distribution suggests a strong alignment between high-quality instruction and high cultural relevance within the dataset. Overall, most of the content of CULTUREINSTRUCT is rated highly in both Instruction Quality and Culture Relevance, emphasizing our data construction pipeline on ensuring both pedagogical effectiveness and cultural sensitivity.

C.3 Dataset Coverage

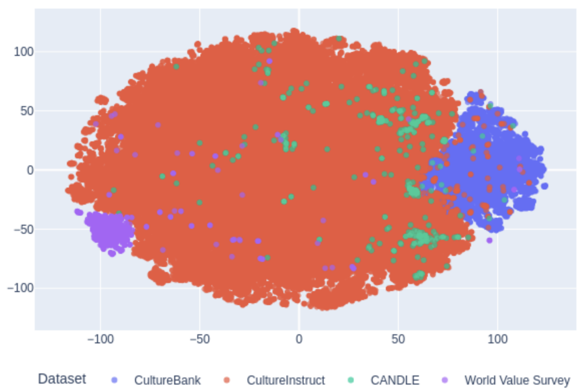

Figure 8: t-SNE scatter plot of CULTUREINSTRUCT and other culture-related dataset, namely CANDLE, CULTUREBANK, WORLD VALUE SURVEY.

Figure 8 illustrates the raw scatter plot of t-SNE projection results. This plot aligns closely with Figure 4 in Section 3.2, where CULTUREINSTRUCT still shows large coverage over other cultural training datasets.

D Additional Experiments

D.1 Evaluation on Standard LLMs Benchmarks

The following benchmarks are chosen for evaluation:

- **HellaSwag** (Zellers et al., 2019) HellaSwag is a benchmark designed to evaluate a model's ability to complete scenarios in a commonsense, plausible manner. It presents a context followed by multiple-choice options, and the model must choose the most likely continuation. HellaSwag is particularly difficult because the incorrect choices are often highly plausible, making it a strong test of a model's commonsense reasoning capabilities.

- **CommonsenseQA** (Talmor et al., 2019) CommonsenseQA is a multiple-choice benchmark designed to test a model's understanding of everyday commonsense knowledge. It presents questions that require reasoning based on basic facts about the world, which are often intuitive to humans but challenging for AI. The questions are crafted from commonsense knowledge sources and are accompanied by five answer choices, only one of which is correct. The benchmark tests the model's ability to reason beyond factual recall, requiring an understanding of how objects, events, and concepts relate in everyday scenarios.

- **GPQA** (Rein et al., 2024) GPQA is a challenging dataset of multiple-choice questions written by domain experts in biology, physics, and chemistry. The questions are high-quality and extremely difficult, and the dataset is split into three sets: **Diamond**, **Extended**, and **Main set**. In our experiments, we use the **Main set** of GPQA for evaluation.

We show the evaluation results of LLAMA-3.1-INSTRUCT and QWEN2-INSTRUCT on standard LLMs benchmark in Table 9. Overall, the performance of LLMs decreases consistently after fine-tuning with CULTUREINSTRUCT. We acknowledge that there exists the catastrophic forgetting problem of these models, as well as there is a domain mismatch between CULTUREINSTRUCT and these benchmarks, hence the performance is lower.

We further tried to fine-tune LLAMA-3.1-INSTRUCT 8B and QWEN2-INSTRUCT 7B with a combined dataset in an attempt to retain some performance on standard benchmarks. Specifically, we combine CULTUREINSTRUCT with a synthesized dataset - MAGPIE (Xu et al., 2024) and fine-tune the models. For LLAMA-3.1-INSTRUCT 8B, we use the `Magpie-Pro-300K-Filtered` subset, while it is `Magpie-Qwen2-Pro-300K-Filtered` subset for QWEN2-INSTRUCT 7B. These sets are chosen so that each model retains some memory of previous fine-tuning. Results of this experiment are shown in Table 10. Overall, fine-tuning with the combined dataset (CULTUREINSTRUCT + MAGPIE) helps to retain some of the performance

Model	Standard LLMs Benchmarks		
	HellaSwag (0-shot) (Accuracy ↑)	CommonsenseQA (Accuracy ↑)	GPQA (main set) (Accuracy ↑)
Parameter Size between 7B and 9B			
LLAMA-3.1-INSTRUCT 8B	75.31	76.00	29.91
+ CULTUREINSTRUCT	74.06	74.61	27.90
QWEN2-INSTRUCT 7B	79.25	80.02	34.15
+ CULTUREINSTRUCT	76.73	73.59	32.28
Parameter Size between 70B and 72B			
LLAMA-3.1-INSTRUCT 70B	84.27	80.10	41.29
+ CULTUREINSTRUCT	83.97	73.05	39.96
QWEN2-INSTRUCT 72B	86.16	88.04	39.06
+ CULTUREINSTRUCT	85.41	87.71	36.38

Table 9: Performance of LLMs on standard LLMs benchmarks that are not related to cultural knowledge.

Model	Cultural Benchmarks (Accuracy ↑)		Standard Benchmarks (Accuracy ↑)		
	CANDLE	CULTURALBENCH	HellaSwag (0-shot)	CommonsenseQA	GPQA
LLAMA-3.1-INSTRUCT 8B	72.20	48.81	75.31	76.00	29.91
+ CULTUREINSTRUCT	79.80	59.52	74.06	74.61	27.90
+ CULTUREINSTRUCT + MAGPIE	74.50	52.79	74.75	74.96	29.24
QWEN2-INSTRUCT 7B	77.80	47.62	79.25	80.02	34.15
+ CULTUREINSTRUCT	88.00	67.46	76.73	73.59	32.28
+ CULTUREINSTRUCT + MAGPIE	81.40	62.05	77.21	78.63	32.31

Table 10: Comparison of smaller LLMs on Cultural and Standard Benchmarks. The MAGPIE subset for LLAMA-3.1-INSTRUCT is `Magpie-Pro-300K-Filtered`, while it is `Magpie-Qwen2-Pro-300K-Filtered` for QWEN2-INSTRUCT.

on standard LLM benchmarks. However, the performance of models on the cultural benchmarks (CANDLE and CULTURALBENCH) is not as good as fine-tuning with only CULTUREINSTRUCT. This is an expected open research problem of LLMs - catastrophic forgetting - and we leave this problem for future works.

```
# Instruction
You need to rate the quality of the question and answer pair based on its clarity,
specificity, and coherence.

The rating scale is as follows:

- very poor: The question and answer pair is unclear, vague, or incoherent. It lacks
essential information and context.
- poor: The question and answer pair is somewhat unclear or lacks important details. It
requires significant clarification.
- average: The question and answer pair is moderately clear and specific. It may require
some additional information for a complete understanding.
- good: The question and answer pair is clear, specific, and mostly well-formed. It
provides sufficient context for understanding the user's intent.
- excellent: The question and answer pair is very clear, specific, and well-articulated.
It contains all the necessary information and context for providing a comprehensive
response.

## Question
{input}

## Answer
{output}

## Response Format
Given the question and answer pair, you first need to give a short assesement,
highlighting the strengths and/or weaknesses of the question and answer pair.
Then, you need to response with a rating from very poor to excellent by filling in the
placeholders in [...]:
{{
    "explanation": "[...]",
    "quality": "[very poor/poor/average/good/excellent]"
}}
```

Figure 9: Prompt template for assessing Instruction Quality.

```
# Instruction
You need to rate the quality of the question and answer pair based on its clarity, specificity,
coherence, and how well it reflects appropriate cultural knowledge. The question and answer
should demonstrate an understanding of cultural context, norms, values, or traditions relevant
to the topic.

The rating scale is as follows:

- very poor: The question and answer pair is unclear, vague, incoherent, or shows a lack of
cultural understanding. It fails to consider essential cultural knowledge, leading to
potentially incorrect or insensitive information.
- poor: The question and answer pair is somewhat unclear or lacks important cultural details. It
reflects little to no cultural context and requires significant improvement in cultural
awareness.
- average: The question and answer pair is moderately clear and specific. It demonstrates some
awareness of cultural knowledge but may lack sufficient context or detail to provide a fully
accurate or nuanced understanding.
- good: The question and answer pair is clear, specific, mostly well-formed, and demonstrates an
appropriate level of cultural understanding. It takes into account important cultural norms or
knowledge relevant to the topic.
- excellent: The question and answer pair is very clear, specific, well-articulated, and
demonstrates deep cultural awareness. It contains all necessary information and reflects a
strong understanding of cultural nuances, context, and relevance.

## Question
{input}

## Answer
{output}

## Response Format
Given the question and answer pair, you first need to give a short assessment, highlighting the
strengths and/or weaknesses of the pair in terms of clarity, specificity, coherence, and how
well it reflects cultural knowledge. Mention any areas where cultural context is either
appropriately addressed or missing.
Then, you need to response with a rating from very poor to excellent by filling in the
placeholders in [...]:
{{
    "explanation": "[...]",
    "quality": "[very poor/poor/average/good/excellent]"
}}
```

Figure 10: Prompt template for assessing Cultural-Relevance.

Domain	**Set of Keywords**
General Cultural Knowledge	Culture, Cultural, Cultural heritage, Tradition, Custom, Folklore, Cultural practice, Ritual, Cultural belief
Art	Arts, Theatre, Cinema, Drama, Painting, Sculpture, Photography, Visual arts, Performing arts, Fine arts, Applied arts
Cuisine	Cuisine, Traditional food, Culinary art, Culinary, Recipe, Gastronomy, Food culture, Food, Ethnic food, Specialty
Cultural Norms	Cultural norm, Social norm, Social practice, Accepted behavior, Traditional practice, Cultural expectation, Social expectation, Social custom, Community standard, Behavioral norm, Cultural standard, Normative behavior, Social rule, Cultural value, Traditional value, Social conduct, Etiquette, Behavioral expectation, Cultural tradition, Tradition, Societal norm, Norms and value, Cultural moral, Social protocol, Normative practice, Social convention, Cultural belief, Ritual practice, Customary behavior, Cultural prescription, Social behavior pattern, Normative social behavior, Cultural code, Social tradition, Traditional social role, Community custom, Cultural conformity, Societal expectation, Cultural practices, Cultural traditions, Social norms, Cultural practices, Customs and rituals, Cultural values, Social behavior, Etiquette and manners, Cultural identity, Cultural diversity, Cultural heritage, Cultural beliefs, Cultural taboos, Social conventions
Festivals	Festival, Celebration, National holiday, Public holiday, Annual event, Ceremony, Ceremonies, National event, Cultural festival, Religious festival, Traditional festival, National holidays, International festival, Harvest festival, Music festival, Film festival, Arts and crafts festival, Food festival, Seasonal festival, Historical festival, Folk festival, Community festival, Festival ritual, Festival custom, Festival tradition, Festival celebration, Festival activities, Festival heritage
History	History, Historical site, Monument, Museum, Archaeology, Ancestry, Genealogy, Historical figure, Cultural landmark, Historical, Ancient history, Medieval history, Modern history, Contemporary history, Historical event, Historical period, Historical figure, Historical movements, Cultural history, Political history, Social history, Economic history, Military history, Diplomatic history, Oral history, Public history, Historiography
Language	Language, Dialect, Linguistic, Idiom, Proverb, Storytelling, Oral tradition, Mythology, Legend, Folktale, Linguistic, Language acquisition, Phonetic, Phonology, Syntax, Semantic, Pragmatic, Morphology, Sociolinguistic, Psycholinguistic, Bilingualism, Multilingualism, Language family, Endangered language, Language preservation, Language evolution, Dialects, Language and culture, Language policy, Translation and interpretation
Literature	Classic literature, Modern literature, Contemporary literature, Literary analysis, Literary criticism, Literary theory, Literary devices, Narrative structure, Fiction, Non-fiction, Poetry, Prose, Literacy
Music	Pop culture, Media, Music history, Music genres, Classical music, Music and culture
Religion	Religion, Spirituality, Belief, Philosophies, Philosophy, Worldview, Religious practice, Sacred text, Deities, Ritual, Religious beliefs, Religious practice, World religions, Monotheism, Polytheism, Spirituality, Sacred text, Religious ritual, Religious symbol, Worship, Prayer, Pilgrimage, Religious ethic, Theology, Comparative religion, Interfaith dialogue, Religious tradition, Religious institution, Clergy, Faith community
Social Life	Social norms, Family structures, Social customs, Daily life, Community life, Social etiquette, Festive events, Traditional clothing, Fashion

Table 11: List of keywords for each domain in the Keyword Filtering step

Culture	Question	Answer
Multiple-choice Question Answering		
China	Which of the following philosophical traditions is the I Ching foundational to in East Asia? A. Buddhism B. Confucianism C. Daoism D. All of the above	The answer is D.
Malaysia	What is a common greeting used in Malaysia during Hari Raya Aidilfitri? A. Selamat Hari Raya B. Maaf Zahir Dan Batin C. Salam Aidilfitri D. Eid Mubarak	C.
Sentence Completion		
Egypt	In Egypt, children typically receive an Eidi, which is a small sum of money to be spent on activities throughout the ________	Eid
Saint Vincent and the Grenadines	Saint Vincent and the Grenadines has a significant diaspora living in ______	the United Kingdom.
Extractive Question Answering		
Canada	Which flag was used as the national flag of Canada until 1965?	The national flag of Canada until 1965 was the Red Ensign, specifically the Canadian Red Ensign. It featured the British Union Jack in the upper left corner and the Canadian coat of arms on a red field.
Fiji	What are the three confederacies into which all Fijian chiefs belong?	Kubuna, Burebasaga, and Tovata.
Textual Entailment		
Indonesia	"The tudong is a style of headscarf, worn as interpretation of the Islamic hijab, prevalent amongst many Muslim women in the Malay-speaking world..." From the context, does it imply that the tudong is a traditional headscarf worn by women in Malaysia and Indonesia? True or False?	True
Ireland	"The president holds office for seven years, and can be ..." - The president of Ireland has significant ceremonial duties but limited executive power.	True
Topic Classification		
New Zealand	Given the following document: "The monarch of New Zealand, who is New Zealand's head of state, is also the Supreme Governor of the Church of England. But the country..." What is the main topic of the document?	Christian politics in New Zealand.
Japan	Given the following document: "Japanese poetry is poetry typical of Japan, or written, spoken..." What is the main subject of the document?	Japanese poetry
Summarization		
America	Given the following document: "This list indexes notable Native American artists from Oklahoma..." What is the primary focus of the document?	The document is a list of notable Native American artists from Oklahoma, Oklahoma Territory, and Indian Territory, detailing their birth or residency in those areas, along with their various artistic disciplines.
Vietnam	Given the following documents: "Chams, or Champa people, Urang Campa, Nguoi Cham, are an Austronesian ethnic group in Southeast Asia and..." What is the historical and cultural significance of the Cham people in Vietnam?	The Cham people, the largest indigenous Muslim group in Vietnam, have a rich history dating back to the 2nd century AD. They are known for their unique language and culture, which have significantly influenced the modern Vietnamese culture.

Table 12: Some examples of each instruction task type in CULTUREINSTRUCT

Document	
It is necessary to remember that each and every culture has its own cultural and social practices. It is understood that in 2 ways, social practice takes place and that is investigation and activity. Social strategies in education include the dynamic use of contact between adults and children to promote growth in children. It is here that literacy activities are seen not only as significant foundations for the regeneration of a society, but also for lifelong learning. Literacy practices are a key are through which instructions are passed on from one generation to another through social identity and social language. This particular paper explores the practices, tenacities, as well as, the discourse of literacy practices that are family based and their connections with the African-American families. A majority of the African-American families are low-income families, thus, their children are usually faced with a higher risk of poor academic achievement when compared to the children that come from white families which are usually either middle class or upper class and this is mainly in the literacy domain. It is acknowledged that there are lifelong consequences of failing to become literate, for example, intergenerational illiteracy, underemployment, and unemployment. It is true to state that the family literacy practices are some of the factors that influence a child's interest in literate behavior. However, these literacy practices vary depending on the contextual, as well as, the cultural variables of each and every family. Some of the literacy practices that have been evidenced to have a positive impact on early literacy among children include parents acting as literacy role models and parent-child book reading and these practices are common among the middle and the upper-class families. There is a clear indication that the literacy skills among the African-American children who mainly come from low-income families have been compromised by the literacy practices of African American families. Family literacy is a social-cultural practice that occurs between the members of a family and acts to promote literacy development, as well as, its use. It can be established that simple literacy practices such as having a large number of books in the house are a literacy practice that influences the literacy development of a child. It is important to note that this is not a common practice among the African American families as a majority of them even lack home libraries. A majority of African American families do not have such and this can be attributed to the fact that a majority of them are single parent families with the single parent having to work long hours so as to meet the financial needs of the family, thus, may lack the time to be close to the children. The other major factor is intergenerational illiteracy. In the early years, only a few African Americans achieved academic success due to a combination of factors, for example, racial discrimination, arrogance, and drugs and substance abuse. Therefore, the African American families lacked literacy role models that the children could look up to and this over a long period of time affected the literacy development of children in African American families. The family literacy practices play a major role in influencing the literacy development of children. However, these family literacy practices are influenced by various contextual and cultural factors. Above are discussed some of the literacy practices that lack in a majority of the African American families and how they have over the time affected the literacy development of the children in those families.	
Generation Results	
Generated Question	**Generated Answer**
What are the consequences of failing to become literate?	Failing to become literate can lead to several lifelong consequences. **Firstly**, it contributes to intergenerational illiteracy, where children from non-literate families may also struggle with literacy. **Secondly**, it can result in underemployment and unemployment, as individuals without literacy skills may face challenges in securing jobs. This can perpetuate the cycle of poverty within families. **Ultimately**, these consequences highlight the importance of literacy not just for individual success but also for societal development.
How do family literacy practices influence children's literacy development?	Family literacy practices significantly impact children's literacy development through various mechanisms. **First**, family literacy practices include activities such as reading with children and providing access to books, which help foster early literacy skills. Research shows that in middle and upper-class families, these practices are common and positively correlate with children's literacy development. **In contrast**, African-American families, particularly those from low-income backgrounds, often lack such practices due to contextual factors like financial constraints and single-parent households. This lack of literacy practices, such as limited access to books and reduced parent-child reading time, adversely affects the literacy development of children. The absence of literacy role models and intergenerational illiteracy further exacerbates the issue. **Therefore**, family literacy practices are crucial in shaping a child's literacy skills, and disparities in these practices can lead to differences in literacy outcomes among different socioeconomic groups.
What are some of the challenges faced by African-American families in supporting their children's literacy development?	African-American families, particularly those from low-income backgrounds, face several challenges in supporting their children's literacy development. **One major challenge** is the lack of resources, such as a sufficient number of books and home libraries, which are crucial for fostering literacy. Many of these families also struggle with limited time for activities like parent-child book reading due to long working hours and single-parenting responsibilities. **Additionally**, historical factors such as intergenerational illiteracy and lack of literacy role models further hinder literacy development. **These factors combined** result in a significant disparity in literacy outcomes between African-American children from low-income families and their peers from middle or upper-class families.

Table 13: Chain-of-Thought Reasoning instructions generated by our tuned BONITO model. **Blue** highlights the beginning of each reasoning step generated by the model.

Question & Answer	Human	LLM	LLM Explanation
Given the following context: "The following outline is provided as an overview of and topical guide to Saint Kitts and Nevis: The Federation of Saint Kitts and Nevis, also known as Saint Christopher and Nevis, is a sovereign federal two-island nation located in the Leeward Islands in the Caribbean Sea. It is the smallest nation in the Americas, in both area and population. The capital city and headquarters of government for the federated state is on the larger island of Saint Kitts. The smaller state of Nevis lies about 2 miles (3 km) southeast of Saint Kitts, across a shallow channel called "The Narrows". Historically, the British dependency of Anguilla was also a part of this union, which was then known collectively as Saint Christopher-Nevis-Anguilla. Saint Kitts and Nevis are geographically part of the Leeward Islands. To the north-northwest lie the islands of Sint Eustatius, Saba, Saint Barthélemy, and Saint-Martin/Sint Maarten. To the east and northeast are Antigua and Barbuda, and to the southeast is the small uninhabited island of Redonda, and the island of Montserrat, which currently has an active volcano (see Soufrière Hills.) Saint Kitts and Nevis were amongst the first islands in the Caribbean to be settled by Europeans. Saint Kitts was home to the first British and French colonies in the Caribbean. Saint Kitts and Nevis is the smallest nation on Earth to ever host a World Cup event; it was one of the host venues of the 2007 Cricket World Cup." **Q:** "What is Saint Kitts and Nevis known for?" **A:** "Saint Kitts and Nevis is a sovereign federal two-island nation in the Caribbean, known for being the smallest nation in the Americas, hosting a World Cup event, and its significant contributions to the music industry, particularly reggae."	Poor	Good	The question and answer pair provides clear and specific information about Saint Kitts and Nevis, including its geographic location, historical significance, and notable achievements. The answer directly addresses the question's intent, highlighting the country's status as the smallest nation in the Americas, its hosting of a World Cup event, and its contributions to the music industry. The answer could benefit from more detail or elaboration on the points raised, but it is clear, concise, and relevant.
Given the following context: "Geography. Maralik is situated at the northwestern foot of Mount Aragats within the Shirak Plain, at a height of 1,720 meters above sea level, at a 24 km road distance south of the provincial centre Gyumri... Demographics. Maralik is home to many historical monuments. The archaeological site at the northeastern vicinity of the town is home to a cyclopean fortress and a settlement, both dating back to the 2nd millennium BC... Transportation. Maralik is located on the M-1 Motorway on the road between Gyumri and Yerevan. The town is also connected with the nearby towns and villages through a network of regional roads... Economy. Maralik was an active industrial town during the Soviet period. The currently-abandoned cotton-spinning factory of Maralik used to have more than 2000 employees at its peak during the Soviet days... Education. Maralik is home to many pre-school kindergartens, 2 public education schools, as well a vocational school which is a stet-owned intermediate technical college, serving the town and its surrounding villages... Sport. Maralik has a sport school run by the municipality. The town is served by a municipal football stadium with a limited seating capacity of 1,000. Football is the most popular sport in Maralik..." **Q:** "What is the main subject of the context?" **A:** "The history, culture, and demographics of Maralik, Armenia."	Poor	Excellent	The question asks for the main subject of the provided context, and the answer succinctly identifies it. Both question and answer are clear, concise, and specific to the task at hand.
Given the following context: "Thomas Sankara at the Marxists Internet Archive." **Q:** "What was the significance of Thomas Sankara during his presidency?" **A:** "Thomas Sankara, known as the 'African Che Guevara', was a visionary leader who fought corruption, promoted free education and healthcare, and worked towards economic independence for Burkina Faso. He was assassinated in 1987."	Very Poor	Good	The question asks for the significance of Thomas Sankara during his presidency, which is clear and specific. The answer provides a good summary of his achievements and...
Given the following context: "In the beginning of the 20th century, Uruguay became the most politically and socially advanced state on the continent. The liberal José Batlle y Ordóñez (in power between 1903 and 1907, then between 1911 and 1915) was the main architect of this transformation; freedom of expression and the press was affirmed, as was that of suffrage. A system of proportional representation was adopted to allow for the representation of minorities. This period also saw the abolition of the death penalty, a fight against administrative corruption, and the introduction of secularism and women's suffrage. On the economic level, Batlle stated that..." **Q:** "What are the main features of Uruguay's political and social transformation in the early 20th century?" **A:** "Uruguay became the most politically and socially advanced state on the continent under the leadership of José Batlle y Ordóñez, who implemented significant reforms including freedom of expression and suffrage, abolition of the death penalty, and nationalization of various sectors."	Good	Average	The answer is clear and specific, but it omits some important details, such as the introduction of women's suffrage and social measures. It also does not provide context on the constitutional and economic aspects of the transformation. The answer could be improved by...
Given the following context: "The right to sit refers to laws or policies granting workers the right to be granted suitable seating at the workplace. Jurisdictions that have enshrined ""right to sit"" laws or policies include Mexico, France, Spain, Argentina, the United Kingdom, Jamaica, South Africa, Eswatini, Cameroon, Tanzania, Uganda, Lesotho, Malaysia, Brazil, Israel, Ireland, Zambia, Guyana, the Indian states of Tamil Nadu and Kerala, the Canadian province of Newfoundland and Labrador, and the British overseas territory of Gibraltar and Montserrat. Almost all states of the United States and Australia, as well as the majority of Canadian provinces passed right to sit legislation for women workers between 1881 and 1917. US states with current right to sit legislation include California, Florida, Massachusetts, Montana, New Jersey, New York, Oregon, Pennsylvania, West Virginia, and Wisconsin. A right to sit provision is included in the International Labour Organization's Hygiene (Commerce and Offices) Convention, 1964; the international treaty being ratified by 52 countries as of 2023. EU-OSHA recommends suitable seating as a best practice..." **Q:** "What is the 'right to sit' and what jurisdictions have enshrined this law?" **A:** "The 'right to sit' refers to laws or policies that grant workers the right to be granted suitable seating at their workplace. Jurisdictions that have enshrined this law include Mexico, France, Spain, Argentina, the United Kingdom, and many others."	Good	Average	The answer is clear and provides a summary of the jurisdictions that have enshrined 'right to sit' laws, but it lacks the detailed and comprehensive information found in the original context, such as specific states, provinces, and countries, as well as international laws and conventions. The answer could be improved by including more specific details from the context.

Table 14: Some examples of instruction quality judgment made by LLAMA-3.1-INSTRUCT 70B and human annotators. The human selections are based on majority voting. Red highlights the problems in the instruction that the LLM Judge has missed. Blue highlights the parts showing the LLM Judge is better than human annotators.