FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models

Shramay Palta
Department of Computer Science
University of Maryland, College Park
spalta@cs.umd.edu

Rachel Rudinger
Department of Computer Science
University of Maryland, College Park
rudinger@umd.edu

Abstract

It is common sense that one should prefer to eat a salad with a fork rather than with a chainsaw. However, for eating a bowl of rice, the choice between a fork and a pair of chopsticks is culturally relative. We introduce FORK, a small, manually-curated set of CommonsenseQA-style questions for probing cultural biases and assumptions present in commonsense reasoning systems, with a specific focus on food-related customs. We test several CommonsenseQA systems on FORK, and while we see high performance on questions about the US culture, the poor performance of these systems on questions about non-US cultures highlights systematic cultural biases aligned with US over non-US cultures.

1 Introduction

Effective communication in natural language requires a shared base of knowledge between interlocutors. While this shared knowledge between communicators may be specific to individuals in a shared situation (e.g., Dhruv and Mei know they are sitting at a table in a restaurant), or to individuals with specialized knowledge (they are discussing backpropagation), some types of knowledge are sufficiently generic to be shared by most people in the world (e.g., *objects fall when they are dropped*). This latter category of *commonsense knowledge* has for decades been a holy grail of research in artificial intelligence and natural language understanding (McCarthy, 1959). If machines are to understand (and produce) human language competently, they must at a bare minimum share this commonsense knowledge with humans.

A question elided by this notion of commonsense knowledge is *who counts as "most people"?* What may appear as universal "common sense" to AI researchers in one cultural context may in fact not be so universal. Early efforts to schematize commonsense knowledge as *scripts*, or stereotyped sequences of events, provide a nice illustration of such unintended cultural biases: the famous "Restaurant script" (Schank and Abelson, 1975) prototypically includes a `LEAVE TIP` event, though tipping is not customary at restaurants in many countries outside the United States.

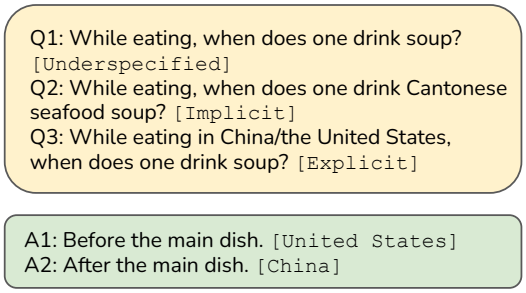


Figure 1: An example from FORK showing an Underspecified, Implicit and Explicit question with the US- and Non-US answer options.

More recent AI research on commonsense knowledge acquisition has relied on crowd sourcing (Regneri et al., 2010; Sap et al., 2019), corpus statistics (Lin and Pantel, 2001; Van Durme and Schubert, 2008), and language modeling (Rudinger et al., 2015; Liu et al., 2022) in place of expert-crafted knowledge. However, each of these methods carries the potential to encode cultural bias into data and models for commonsense reasoning, whether through the implicit cultural perspectives of corpus texts, crowd source workers, or AI researchers themselves.

In this work, we seek to investigate cultural biases or assumptions present in commonsense reasoning systems. Culture, like commonsense knowledge, is vast. By one definition,[1] *culture* "encompasses the social behaviour and norms found in human societies, as well as the knowledge, beliefs, arts, laws, customs, capabilities, and habits of the individuals in these groups." From the social sciences, Kendall (2015) defines culture as encom-

[1] https://en.wikipedia.org/wiki/Culture

Findings of the Association for Computational Linguistics: ACL 2023, pages 9952–9962
July 9-14, 2023

passing both material as well as non-material aspects, such as beliefs and linguistic practices. To limit the scope of our investigation, however, we focus on a single topic common to all human cultures but widely varying across them: food.

We introduce FORK (**F**ood **OR**iented cultural commonsense **K**nowledge), a manually-curated set of CommonsenseQA-style (Talmor et al., 2019) test questions for probing culinary cultural biases and assumptions present in commonsense reasoning systems. For the purpose of this work, we say that a commonsense question-answering system is *culturally biased* if (1) in response to questions with culturally-dependent answers, it exhibits systematic preference for answers consistent with one cultural setting over others; or (2) for questions with explicit cultural contexts, it exhibits systematically higher accuracy for some cultural contexts over others. Figure 1 contains an example of three interrelated test questions in FORK we use to detect cultural bias. For Q1, a model that prefers A1 to A2 exhibits cultural bias in favor of the United States (US) over China. While there exists no tidy mapping between human cultures and countries, in this work, we use countries as a coarse-grained proxy for culture (see: § 7).

FORK contains questions pertaining to the food and culinary cultures of the US, China, Japan, and India with questions spanning topics of restaurant tipping, eating utensils, and other culinary customs. We test multiple encoder-based CommonsenseQA models on FORK, demonstrating systematic cultural biases favoring the US over non-US countries.

To summarize, our contributions are:

1. FORK: a "bite-sized" manually curated test set of 184 CommonsenseQA-style questions which can be used for probing culinary cultural biases and assumptions in commonsense reasoning systems.
2. A systematic evaluation of several encoder based models on FORK to demonstrate systematic cultural assumptions aligned with US over non-US cultures.

2 Dataset

Since FORK aims to test the culinary cultural specificity of commonsense reasoning models, we choose the format to be along the lines of Commonsense QA (Talmor et al., 2019). Each question in FORK has two options, only one of which is correct. One of the options pertains to the US culture, while the other to non-US. The questions are manually written by the first author of this paper. The source of content used to formulate the questions is information gathered from Google searches, blog posts, traveler guides, etc. Upon publication, we will release FORK publicly.

There are three types of questions in FORK:

- **Underspecified**: The question asked is about culinary customs and practices of no particular country or culture, and we hypothesize that English models will default to a US-centric interpretation in such a case. (See Fig. 1, Q1.)
- **Implicit**: The question asked is about culinary customs and practices in context of a particular country but no country is mentioned explicitly in that question. Rather, the cultural setting is established implicitly with context cues. (See: Fig. 1, Q2.)
- **Explicit**: The question asked is about culinary customs and practices in context of a particular country and that country (or well-known city therein) is explicitly mentioned in that question. (See: Fig. 1, Q3.)

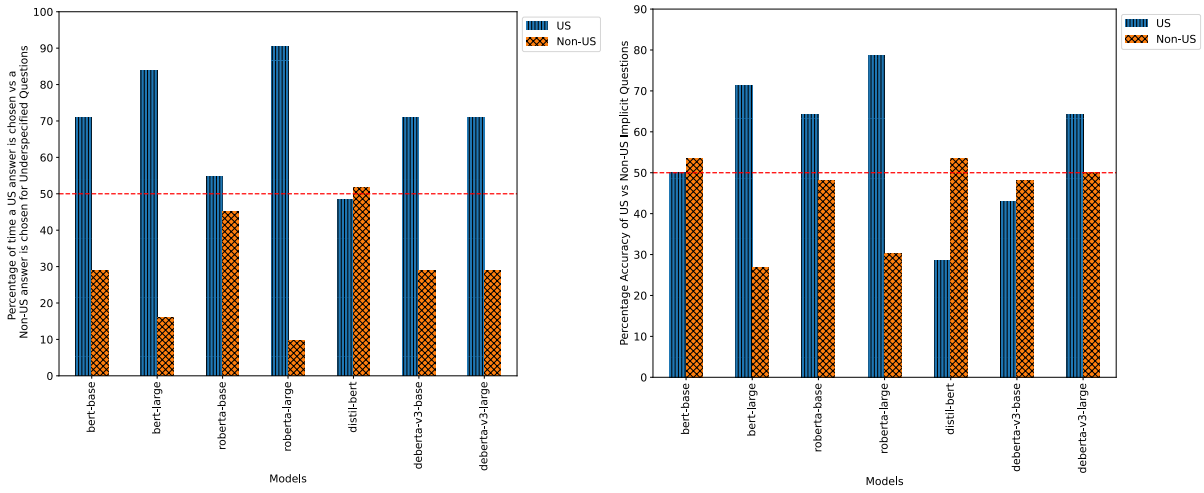
We assign a theme to each question, and questions in FORK span over three distinct culinary themes: *eating utensils*, *tipping* and *general custom/culture*. We also tag each Underspecified question-answer pair, and each implicit and explicit question, with a corresponding country. We present a brief overview of the distribution in Table 1, and a full demographic distribution in Table 3.

Country	Underspecified	Implicit	Explicit	Total (by country)
US	31	14	13	58
Non-US	0	56	70	126
Total (by type)	31	70	83	184

Table 1: Number of questions in FORK from different types and countries.

It is important to note that these country labels should not be construed as *exclusive* of other cultures or countries that may share the relevant attribute. Cultural customs and countries have a many-to-many relation, and our labels are intended to highlight particular points of contrast between the US and other countries. What we measure as US-oriented cultural bias could also be construed as, e.g., Canada-oriented cultural bias only to the extent that US-labeled questions are also applicable to Canada.

The questions in FORK can either be a single



(a) (b)

Figure 2: Results for Underspecified and Implicit Questions: a) Percentage times a US answer is chosen vs a non-US answer is chosen for Underspecified questions. b) Percentage accuracy for US and non-US Implicit questions.

sentence or a two sentence question. The questions consisting of two sentences help to provide context in case of *Implicit* and *Explicit* questions. We also follow a template-based approach where a question about the same theme is asked multiple times, varying only by e.g. the name of a dish, city, country, etc.

In total, FORK consists of 184 manually curated questions, with 91.84% of questions pertaining to China, Japan, India and the US, and a small number of additional questions for other countries. Researching and writing questions was a slow manual process, so we chose to focus on producing more questions for fewer countries, to yield more robust results.

2.1 Validation Study

In order to ensure that the manually curated questions are valid for probing culinary cultural differences, we conduct a validation study with six annotators, two each from China, the USA and India. This pool of annotators comprised of five graduate students, and one professor. We ask the annotators to answer questions in FORK and present statistics in Table 2.

Country	Cohen's Kappa	Raw Agreement
US	1.0	100%
China	0.93	96.96%
India	0.52	76.47%

Table 2: Results from the validation study to attest the quality of questions in FORK

For US, both annotators disagreed on the same question, while for India, the difference was on questions pertaining on tipping. Feedback from annotators observed that the tipping culture varied across the country. For China, the human annotators noted that some practices were untrue for the regions they were from, but true for other regions in China.

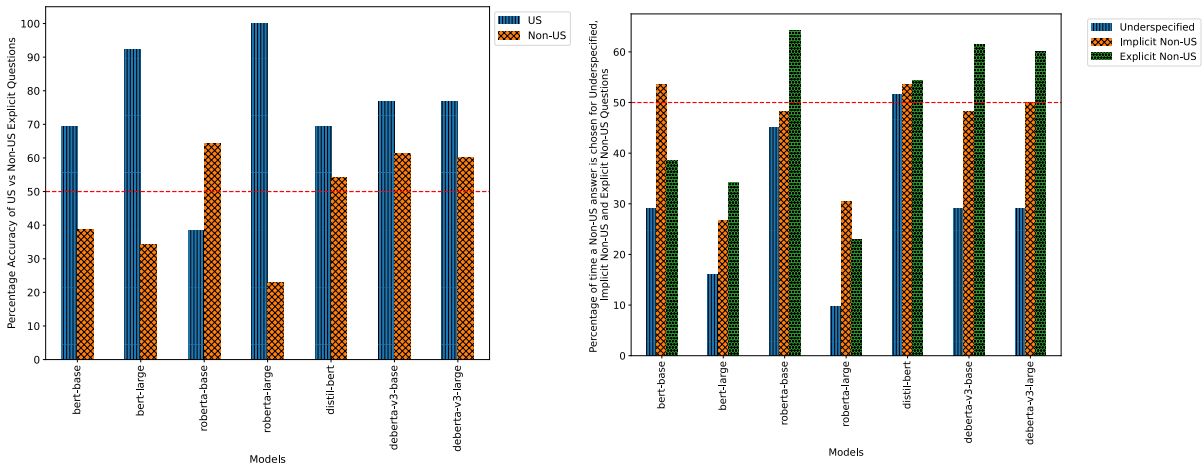
Additionally, the differences in customs and practices within the same country reiterate our note above that cultural customs and countries have a *many-to-many* relation. We have used *country* as a proxy variable for culture because there are no clear distinct boundaries across cultures, and using this proxy boundary allows to probe differences at a US vs non-US level. This highlights a need for future work to investigate cultural differences *within* a country, based on regional or other demographic dimensions.

3 Experiments

We summarize our experimental set up, models, and the evaluation strategy used in this work.

3.1 Experimental Setup

In this work, we test seven encoder-based models on FORK and report their performance. We test two variants of BERT (Devlin et al., 2019): `bert-base` and `bert-large`, two variants of RoBERTa (Liu et al., 2019): `roberta-base` and `roberta-large`, DistilBERT (Sanh et al., 2019), and two variants of DeBERTaV3 (He et al., 2021): `deberta-v3-base` and `deberta-v3-large`.



(a) (b)

Figure 3: a) Percentage Accuracy for US and non-US Explicit questions. b) Percentage times a non-US answer is chosen for Underspecified, Implicit non-US and Explicit non-US questions.

All models are finetuned on the CommonsenseQA `train` fold for 3 epochs. We run a grid search for the hyper-parameters and report them in Appendix A.2.

3.2 Evaluation Strategy

We evaluate the culinary cultural contingency of the models tested as follows. For the questions tagged as *Underspecified*, we look at the number of times a "US" answer is chosen over a "non-US" answer. Here, "US" answer refers to an answer that would be appropriate or likely in the US context, and "non-US" answer refers to an answer that is more appropriate for a country on context outside the US. For *Implicit* and *Explicit* questions, we take a look at the responses for both US and non-US questions, and the percentage accuracy for US vs non-US answers.

Additionally, we also compare the number of times a non-US answer is chosen for *Underspecified*, non-US *Implicit* and non-US *Explicit* questions to better determine the bias between US and non-US cultures.

4 Results and Discussion

Figure 2a shows the percentage of time when a US answer is chosen over a non-US answer for *Underspecified* questions. We observe that `roberta-large`, and `bert-large` report the top two percentages of choosing a US answer over a non-US answer with values of 90.32% and 83.87% respectively. Fig 2a shows that all models, except for `DistilBERT`, preferred US answers over non-US answers for a majority of *Underspecified* questions.

Figure 2b shows the percentage accuracy for US and non-US *Implicit* questions. We observe that `roberta-large` and `bert-large` report the top two accuracies of 78.57% and 71.42% respectively, when answering US *Implicit* Questions. In contrast, for non-US *Implicit* questions, only two models, `DistilBert` and `bert-base` cross the 50% accuracy mark, with `bert-large` having the lowest accuracy of 26.78%.

Figure 3a shows the percentage accuracy for US and non-US *Explicit* questions. Here, `roberta-large`, and `bert-large` report the top two accuracies of 100% and 92.30% respectively when answering US *Explicit* Questions. In contrast, for non-US *Explicit* questions, `roberta-base` reports the best accuracy at 64.28% while `roberta-large` performs the worst, achieving 22.85%.

Figure 3b shows the percentage times a non-US answer is chosen for *Underspecified*, non-US *Implicit* and non-US *Explicit* questions. For *Underspecified* questions, only `DistilBert` crosses the 50% mark, with 51.62% accuracy. The performance for non-US *Implicit* and non-US *Explicit* questions has been discussed above. We report all the model accuracies on FORK in Tables 4 and 5 in Appendix A.2.

In addition to aggregating US versus non-US results, we break down accuracy results for China, India, and Japan for Implicit and Explicit questions in Table 6 in Appendix A.2.

We observe that *(China, Explicit)* and *(China, Implicit)* questions have the lowest average accuracy across models, at 36.57% and 41.80%, respectively. The best performance is reported for *(USA, Explicit)* at 74.72%.

4.1 Statistical Significance of Results

In order to make sure that our findings our statistically significant, despite the small number of questions in FORK, we conduct several statistical significance tests.

For Underspecified questions, we conduct the binomial test for all 7 model prediction results separately. Only `roberta-base` and `DistilBert` report a p-value greater than 0.05 in this setting.

For Implicit and Explicit questions, we conduct the chi-squared test on all 7 model prediction results separately to determine the statistical significance of our findings. For Implicit questions, only `bert-base` and `deberta-v3-base` report a p-value greater than 0.05 respectively. None of the models reported a p-value greater than 0.05 for Explicit questions.

5 Related Works

A growing body of work aims to detect social biases in NLP models with respect to demographic attributes like gender and race (Rudinger et al., 2018; Zhao et al., 2018; Nangia et al., 2020; Li et al., 2020; Sap et al., 2020). More recent is the growing attention towards cultural biases in NLP and AI technology at large. Hershcovich et al. (2022) propose a framework that allows one to understand the challenges of cultural diversity in NLP applications. Wikipedia has been shown to embed latent cultural biases (Callahan and Herring, 2011) and Tian et al. (2021) propose a methodology to develop culturally aware models for English, Chinese and Japanese using distributional perspectives on controversial topics from Wikipedia across these languages. Acharya et al. (2020) explore cultural biases, but along the rituals like birth, coming of age, marriage etc. in the US and in India. Chen and Henning (1985) investigate cultural bias in language proficiency tests and identify items of bias against non-native English speakers. To the best of our knowledge, this is the first work to analyze cultural bias in commonsense reasoning from the angle of culinary customs.

6 Conclusion

We have introduced FORK, a dataset to measure culinary cultural bias in commonsense models. Confirming our hypothesis, we find that models default to US cultural contexts in underspecified questions, and perform markedly better on implicit and explicit questions about US culture than non-US. A likely source of bias is the English, US-produced texts which models are pretrained on. We believe the results support our hypothesis that English Language Models LMs trained on texts (many of which are produced for a US audience) would reflect US (or broadly Western) cultural assumptions. We hypothesize that the Underspecified setting had the lowest "accuracy" for non-US countries because the experimental design forced the model to choose between US and non-US interpretations of the same question. For Implicit and Explicit settings, we speculate that the non-US accuracy is generally higher for Explicit than Implicit because the former made it easier for models to determine the cultural setting.

Potential mitigation techniques to eliminate such biases may involve better curation of training data, training separate models for different cultural contexts, training models to better recognize cultural cues or ask for clarification in ambiguous settings, among many other possibilities. We believe this is an open research question, and we hope this paper will inspire future research to address it.

The topic of cultural bias is vast, and we choose a narrow scope to avoid biting off more than we can chew. Future work will explore strategies for cultural awareness of commonsense models, analysis of cultural assumptions in non-English models, and analysis of other aspects of culture beyond the culinary.

7 Limitations

The term *culture* has many meanings, and before attempting to incorporate commonsense with culture, one needs to establish a well defined definition and boundary along which test cases and examples would be constructed. By focusing exclusively on food and culinary customs, we have greatly restricted our domain of inquiry. However, culinary topics are universal, and span multiple domains of common sense reasoning (physical, interpersonal, societal). Nonetheless, we hope this work will inspire future work to investigate cultural bias along many axes beyond the culinary.

Incomplete representation of all cultures: There are limitations with using countries as a proxy for culture. As noted in § 2, mappings between cultures and countries are many-to-many, not one-to-one. The majority of questions in our test set FORK focus on culinary cultures and customs of only a few countries, and we do not expect the results to generalize to all the countries of the world. We choose to focus only on one topic and a small number of countries so that we may initiate research on this broad, challenging problem with a narrower, more well-defined task. We selected these cultures based on the cultural backgrounds of the authors and authors' colleagues who were available to provide direct feedback on/validate the data. We hope this work paves the way for follow-up work investigating a broader set of cultures.

Small Annotator Pool: The validation study in § 2.1 is done on a small pool of annotators from a few countries represented in FORK. While the study gave useful feedback about the dataset and question quality, a larger and more diverse set of annotators would reflect a broader range of perspectives within each country, and further reduce the potential for biases or inaccuracies in our data.

8 Ethics Statement

Our paper has demonstrated systematic cultural biases in commonsense reasoning models' understanding of culinary scenarios. While our end goal is to develop methods of evaluating cultural biases in models between the US and other countries, we acknowledge a number of risks involved in this endeavor. In particular, we note that any attempts to define, characterize, or delineate different cultures, particularly those to which the authors do not belong, creates a potential for oversimplifying the representations of those cultures and failing to represent minority populations therein. To mitigate this, we had a small number of annotators from the US, China, and India validate the questions, but these annotators do not represent the full diversity of each of these countries. We also caution that, while this dataset may be used to demonstrate the *presence* of cultural biases in commonsense reasoning models, it cannot be used to prove the absence thereof.

9 Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback on this paper. Additionally, we would like to thank Antoine Bosselut, Linda Zou, Peter Rankel, Abhilasha Sancheti, Haozhe An, Elijah Rippeth, Rupak Sarkar, Ishani Mondal, and Rebecca Knowles for their helpful comments and suggestions.

References

Anurag Acharya, Kartik Talamadupula, and Mark A Finlayson. 2020. Towards an atlas of cultural commonsense for machine reasoning. *arXiv preprint arXiv:2009.05664*.

Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, 62(10):1899–1915.

Zheng Chen and Grant Henning. 1985. Linguistic and cultural bias in language proficiency tests. *Language Testing*, 2(2):155–163.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Diana Kendall. 2015. *Sociology in our times: The essentials*, 10 edition. CENGAGE Learning Custom Publishing, Mason, OH.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Dekang Lin and Patrick Pantel. 2001. Dirt @sbt@discovery of inference rules from text. In *Proceedings of the Seventh ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, KDD '01, page 323–328, New York, NY, USA. Association for Computing Machinery.

Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

John McCarthy. 1959. Programs with common sense.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 979–988, Uppsala, Sweden. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Rachel Rudinger, Pushpendre Rastogi, Francis Ferraro, and Benjamin Van Durme. 2015. Script induction as language modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1681–1686, Lisbon, Portugal. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Roger C Schank and Robert P Abelson. 1975. Scripts, plans, and knowledge. In *IJCAI*, volume 75, pages 151–157.

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.

Yufei Tian, Tuhin Chakrabarty, Fred Morstatter, and Nanyun Peng. 2021. Identifying distributional perspectives from colingual groups. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 178–190, Online. Association for Computational Linguistics.

Benjamin Van Durme and Lenhart Schubert. 2008. Open knowledge extraction through compositional language processing. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 239–254. College Publications.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

A.1 Details about FORK

We present the full details about our test bed FORK. FORK has a total of 184 questions with 169 (91.84%) questions talking about the culinary culture in the US, China, Japan and India. We present the full demographic distribution of questions in Table 3

Country	Underspecified	Implicit	Explicit	Total (by country)
US	31	14	13	58
China	0	41	25	66
Japan	0	6	22	28
India	0	6	11	17
UAE	0	0	4	4
France	0	1	1	2
Germany	0	0	2	2
Italy	0	1	1	2
South Korea	0	1	1	2
Saudi Arabia	0	0	2	2
Kazakhstan	0	0	1	1
Total (by type)	31	70	83	184

Table 3: Demographic distribution of questions in FORK

A.2 Details about Model Parameters and Performance on FORK

Model	Underspecified	
	US	Non-US
bert-base	70.96	29.04
bert-large	83.87	16.13
roberta-base	54.83	45.17
roberta-large	**90.32**	9.68
distil-bert	48.38	51.62
deberta-v3-base	70.96	29.04
deberta-v3-large	70.96	29.04

Table 4: Percentage times a US answer is chosen vs a non-US answer is chosen for Underspecified questions for all models.

Model	Type			
	Implicit		Explicit	
	US	Non-US	US	Non-US
bert-base	50.0	**53.57**	69.23	38.57
bert-large	71.42	26.78	92.30	34.28
roberta-base	64.28	48.21	38.46	**64.28**
roberta-large	**78.57**	30.35	**100.0**	22.85
distil-bert	28.57	**53.57**	69.23	54.28
deberta-v3-base	42.85	48.21	76.92	61.42
deberta-v3-large	64.28	50.0	76.92	60.0

Table 5: Percentage accuracies of the models for Implicit and Explicit questions for both US and non-US countries.

We fine-tune each of the models mentioned earlier on the CommonsenseQA `train` set with a grid search for hyper-parameters [batch size = 16 (10 for the DeBERTaV3 models), learning rate = $\{\{2, 3, 4, 5, 6, 7\}e^{-4,-5,-6,-7}\}$, epoch = $\{3, 5, 10\}$]. Training for 3 epochs gives us the best performance on the CommonsenseQA `validation` fold.

For the BERT(Devlin et al., 2019) models, we end up using a learning rate of $4e^{-5}$, and $6e^{-5}$ for `bert-base` and `bert-large` respectively. For RoBERTA models (Liu et al., 2019), we use $3e^{-5}$, and $7e^{-6}$ for `roberta-base` and `roberta-large` respectively. For DistilBERT (Sanh et al., 2019), and two variants of DeBERTaV3 (He et al., 2021): `deberta-v3-base` and `deberta-v3-large` a learning rate of $3e^{-5}$, $2e^{-5}$, and $7e^{-6}$ gives us the best performance respectively. We used the RTX A6000 GPU and finetuning the models took approximately 4 hours. Performance statistics of all the fine-tuned models on FORK are reported in tables 4, 5 and 6.

Table 4 shows the percentage times each model chooses a US answer and non-US answer when answering `Underspecified` questions from FORK. `roberta-large` chooses a US answer 90.32% of the time while all other models, except DistilBERT, choose US answers more than 50% of the time. Overall, we can observe that the models tend to choose a US answer more as compared to a non-US answer.

Table 5 shows the percentage accuracy of each model when answering `Explicit` and `Implicit` questions from FORK. For `Explicit` questions, we can see that one model achieves a 100% while answering questions about the US, while the highest accuracy for non-US questions is at 64.28%. Similarly for `Implicit` questions, the highest accuracy for US questions is 78.57% while it is at 53.57% for non US questions.

Table 6 shows statistics for a *(Country, Type)* pair that has at least 10 questions in FORK and the performance of the fine-tuned models. We observe that the average accuracy for *(*US*, Explicit)* is the highest followed by *(*US*, Underspecified)* at 74.72% and 70.04% respectively. In contrast, only *(India, Explicit)* gets an average accuracy higher than 50%.

All these observations clearly highlight the existence of systematic cultural assumptions aligned with US over non-US countries.

A.3 Licenses

We have used BERT, RoBERTa, DistilBERT and DeBERTaV3 in this work. All these models use Apache License Version 2.0.[2] The CommonsenseQA Dataset is under the MIT License.[3] We are granted permission to use and modify these models for our experiments as per the terms of these licenses.

[2] https://www.apache.org/licenses/LICENSE-2.0
[3] https://opensource.org/licenses/MIT

Country-Type	Total Questions	Model							Average Accuracy
		bert-base	bert-large	roberta-base	roberta-large	distil-bert	deberta-v3-base	deberta-v3-large	
China, Implicit	41	58.53	21.95	48.78	21.95	56.09	41.46	43.90	41.80
USA, Underspecified	31	70.96	83.87	54.83	90.32	48.38	70.96	70.96	70.04
China, Explicit	25	56.0	8.0	56.0	16.0	40.0	40.0	40.0	36.57
Japan, Explicit	22	40.90	22.72	81.81	22.72	68.18	59.09	59.09	50.64
USA, Implicit	14	50	71.42	64.28	78.57	28.57	42.85	64.28	57.13
USA, Explicit	13	69.23	92.30	38.46	100.0	69.23	76.92	76.92	**74.72**
India, Explicit	11	18.18	72.72	54.54	36.36	63.63	100.0	90.90	62.33

Table 6: Total number of questions for each (Country, Type) pair and percentage accuracy for each model.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7: Limitations

☑ A2. Did you discuss any potential risks of your work?
Section 8: Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, Abstract and Section 1: Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2: Dataset and Section 3: Experiments

☑ B1. Did you cite the creators of artifacts you used?
Section 2: Dataset and Section 3: Experiments

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A.3

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A.3. We haven't mentioned license information for our artifact. We will do that when we release the full artifact.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4: Results and Discussion and Appendix A.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2: Dataset and Appendix A.1

C ☑ Did you run computational experiments?

Section 3: Experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3: Experiments and Appendix A.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4: Results and Discussion and Appendix A.2

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 2: Dataset

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No, not included in the paper. The task was to choose between two options for a question.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No, not required for this study. Annotators were colleagues of the authors.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not Applicable. No PII data collected.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not required.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Yes, Section 2 Dataset