Salad-Bowl-LLM: Multi-Culture LLMs by In-Context Demonstrations from Diverse Cultures

Dongkwan Kim
KAIST
dongkwan.kim@kaist.ac.kr

Junho Myung
KAIST
junho00211@kaist.ac.kr

Alice Oh
KAIST
alice.oh@kaist.edu

Abstract

Large Language Models (LLMs) have shown proficiency in various tasks but often struggle to capture cultural knowledge, especially for underrepresented regions. To adapt LLMs to diverse cultures, we explore the power of in-context learning (ICL), where models can leverage contextual demonstrations. Specifically, we investigate the effect of the same, different (i.e., cross-cultural), or flawed in-context demonstrations on a cultural question-answering task across 16 cultures. Our findings show that demonstrations from the same culture generally enhance performance, and cross-cultural demonstrations sometimes outperform those from the same culture. However, incorrect cross-culture demonstrations can substantially decrease performance. These results suggest that knowledge of well-known cultures can potentially enhance the models' understanding of marginalized ones. We leave how to choose which culture's demonstrations for future work to reflect better the diversity of cultures within LLMs.

1 Introduction and Related Work

Large Language Models (LLMs) have demonstrated strong generalization across domains and tasks. However, their ability to represent diverse cultures remains lacking, particularly for underrepresented ones in the primary training sources [1, 9]. This limitation raises concerns about the equity and inclusivity of LLMs, as culture-specific common sense and social norms from marginalized cultures are often overlooked [15].

There are various ways, such as fine-tuning, to adapt language models to incorporate information from little-known cultures. Previous studies have demonstrated that fine-tuning LLMs on culture-specific datasets improves performance on downstream cultural tasks [5, 12, 14, 18]. However, fine-tuning LLMs on additional data is highly resource-intensive and often impossible. Other methods of cultural alignment involve prompting but are mostly limited to providing personas or additional socio-cultural background [1, 13, 8].

An alternative approach is in-context learning (ICL). This simple yet powerful method provides the model with task-specific examples, or "demonstrations," directly in the input without modifying the model's parameters [2]. This allows the LLM to learn dynamically from these examples, making ICL attractive for tasks like cultural knowledge transfer.

In this paper, we examine how well ICL can adapt models to cultural question-answering tasks by leveraging contextual demonstrations that represent diverse cultural backgrounds. Specifically, we employ various in-context demonstration strategies inspired by previous studies [6, 11, 7, 16] to conduct ICL with demonstrations from the same, different, and anonymized cultures and mislabeled answers. We evaluate these strategies on a cultural QA benchmark across 16 cultures [9] with two widely used LLMs, GPT-4o (2024-05-13) and GPT-3.5 (turbo-1106).

Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024.

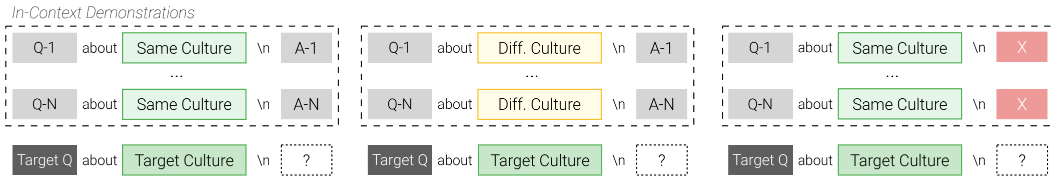

Figure 1: Settings for In-Context Demonstrations. **Left (Same):** Demonstrations align with the target's culture. **Center (Different):** Demonstrations are from different cultures. **Right (No or Wrong Labels):** Demonstrations are from the same culture, but answers are absent or incorrect.

Our experiments show that providing in-context demonstrations from the same culture generally improves the model's performance in cultural tasks, though this is not always guaranteed. Interestingly, cross-cultural demonstrations sometimes result in even higher performance compared to those from the same culture, suggesting that knowledge from certain cultures may help models generalize or enrich their understanding of others. However, the model's performance significantly decreases when the wrong cross-cultural demonstrations are used. Contrary to findings from previous ICL studies [7], we observe that model performance drops sharply when random labels are given in the demonstration. Based on our findings, we open one direction for future work, particularly how to select the most appropriate cultural demonstrations to enhance multicultural LLMs.

The main contributions of our paper are as follows. First, we introduce a novel exploration of how in-context learning (ICL) performs across diverse cultural settings (§2). Second, we conduct comprehensive experiments to assess the impact of same-culture and cross-culture demonstrations, across 16 cultures (§3 and §4). Third, we discuss the implications and future directions in building more inclusive multi-culture LLMs (§5 and §6).

2 In-Context Demonstrations for Cultural Tasks

This paper employs the standard setting of in-context learning (ICL), wherein a model is provided with a series of input-output demonstrations (i.e., question-answer pairs) before answering a target question. We hypothesize that cultural knowledge embedded in these demonstrations can influence the model's performance on a given cultural question.

To investigate this hypothesis, we design three settings to configure in-context cultures and answers as illustrated in Figure 1. In addition, we can conduct experiments using anonymized demonstrations. Prompt examples are provided in the Appendix A.

The Same Culture We feed demonstrations that match the target question's culture. For example, if the target question is about the United States, all contextual demonstrations are also related to the United States. This measures the LLM's ability to utilize relevant cultural knowledge from demonstrations.

Different Cultures The demonstrations come from cultures different from that of the target question. For instance, a target question on the United States can be accompanied by demonstrations of Mexican culture. This setting tests the LLM's ability to transfer knowledge across cultural contexts.

Wrong Labels The model is provided with demonstrations from the same culture as the target question, similar to the 'Same Culture' setting, but the answers in the demonstrations are incorrect. This allows us to test LLMs when exposed to cultural cues but incorrect information. Previous ICL research suggests that the model's performance is not significantly affected by such flawed demonstrations [7].

Culture Anonymization This setting replaces culture names in the demonstrations with random hashes and is applied exclusively to the 'Same Culture' setting. We can explore whether the model's performance is truly based on the cultural knowledge provided by the demonstrations or if it relies on explicitly provided names. For example, "in the United States" is anonymized to "in ABCBF." This helps isolate whether cultural understanding stems from the content of the demonstrations themselves or from the explicit labeling of cultural context.

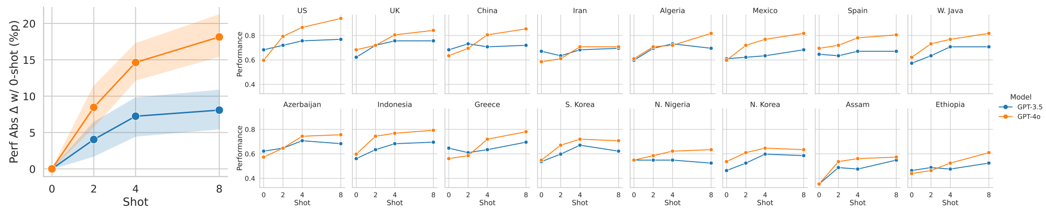

Figure 2: Performance where cultures are *anonymized*. **Left:** Average absolute performance difference with 0-shot across cultures. **Right:** Performance by the number of demonstrations per culture. Shaded bands indicate 95% confidence intervals.

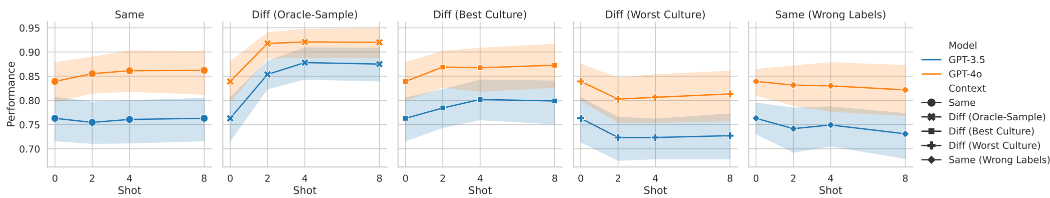

Figure 3: Performance by the number of in-context demonstrations (shots) with five demonstrations: same culture, different cultures (Oracle-Sample, Best Culture, and Worst Culture), and same culture with wrong labels. In the oracle-sample setting, we select a culture for ICL per each question to which the LLM correctly responds. For best and worst cultures, we choose a single culture for ICL per the target culture where the model is most or least accurate. Shaded bands indicate 95% confidence intervals.

3 Experiments

Dataset For the experiment, we exclusively use the multiple-choice questions from BLEnD [9], a cultural QA dataset covering 16 countries and regions. The dataset includes questions that reflect everyday life on six key topics: *food*, *sports*, *family*, *education*, *holidays/celebrations/leisure*, and *work-life*. We select BLEnD as it covers diverse regions—including underrepresented countries—with a balanced set of questions for each country unlike other cultural QA datasets [17, 4, 3]. In addition, BLEnD provides answers for all 16 countries to the same set of questions, making it ideal for ICL. Specifically, we randomly sample multiple-choice questions from BLEnD to create a unique set of 82 questions per country, resulting in a total of 1,312 questions.

Experimental Settings We conduct experiments using two well-known LLMs, GPT-4o (2024-05-13) and GPT-3.5 (turbo-1106) [1], with varying numbers of in-context demonstrations (0, 2, 4, and 8). For demonstrations, similar questions to the target test question are selected. After anonymizing the cultural identifier, we choose questions with embeddings close to the test question, encoded by a pretrained Sentence-Transformer [10] [2]

4 Results and Discussions

In-Context Learning on Anonymized Demonstrations In Figure 2, we observe that increasing the number of demonstrations improves model performance across different cultures. LLMs can infer cultural knowledge from context alone, even when cultural identifiers are removed. Furthermore, we note a lower performance for underrepresented cultures when there is no demonstration (i.e., 0-shot). For example, unlike the United States (0.60 – 0.68) or China (0.63 – 0.68), the 0-shot performance of Assam (0.35) or Ethiopia (0.44 – 0.46) is substantially low. Since only random guesses are possible in an anonymized 0-shot, this result highlights an inherent bias in LLMs toward representative cultures.

In-Context Learning on Demonstrations from Same and Different Cultures We plot the results on the same and different demonstrations in Figure 3 and Figure 4 (per culture). The different

[1]https://platform.openai.com/docs/models
[2]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

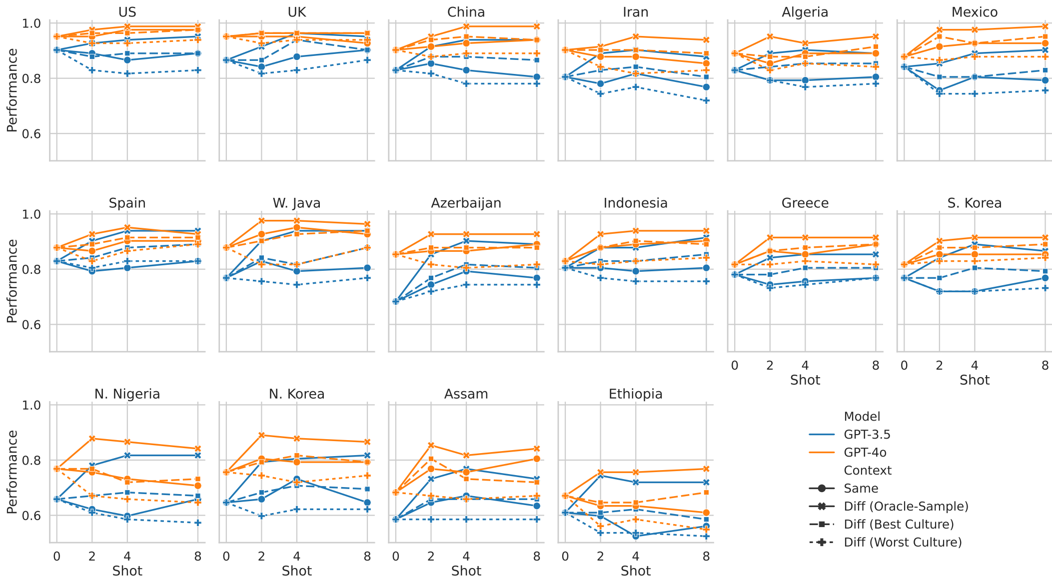

Figure 4: Performance by the number of in-context demonstrations per culture, comparing same to different culture demonstrations with highest and lowest performances.

cultures with the highest and lowest performance are only displayed in these figures. While adding in-context samples from the same culture improves performance in GPT-4o, GPT-3.5 shows minor changes compared to the 0-shot performance. GPT-3.5 has performance gains in some cultures (e.g., West Java, Azerbaijan, North Korea) and losses in others (e.g., Iran, Northern Nigeria, Ethiopia). Performance improvement does not correlate significantly with whether a culture is underrepresented. For example, performance increases by shots consistently in North Korea and Assam but decreases in Northern Nigeria and Ethiopia.

Surprisingly, demonstrations from different cultures sometimes contribute more significantly to performance improvements than the same culture, particularly in cases like China, Spain, and South Korea. In particular, selecting the appropriate culture for each sample (i.e., Oracle-Sample) can yield better performance gains than using the same culture across all samples (i.e., Best Culture).

However, not all different cultures aid in performance enhancement—some even degrade it more than the 0-shot condition across all experiments. This nuanced interaction suggests that while cultural demonstrations can enhance cultural understanding, selecting appropriate cultural contexts remains critical for optimizing performance.

Effects of Absent and Incorrect Labels in In-Context Demonstrations We also find that incorrect labels, known to impact ICL performance barely [7], degrade performance in our culture-related benchmarks. This decline in performance occurs regardless of models. The performance decrease in wrong labels is not mitigated as the number of shots increases.

Performance Increase Rate and Cultural Proximity We analyze the correlation between performance increment by shots and cultural proximity. For each experiment (test and context), we fit linear least squares to the performance by the number of demonstrations. Figure 5 (Left) presents the slope in the fitted equation as a heat map, representing the performance increase rate by shots. We compute the correlation between these slopes and cultural proximity, defined as the average number of common lemmas in each answer (Figure 2 in Myung et al. [9]). On the right side of Figure 5, GPT-4o shows a modest positive correlation (0.37), indicating it better captures cultural relationships inferred from common lemmas, while GPT-3.5 shows almost no correlation (0.08). This suggests that GPT-4o is more sensitive to cultural proximity in demonstrations.

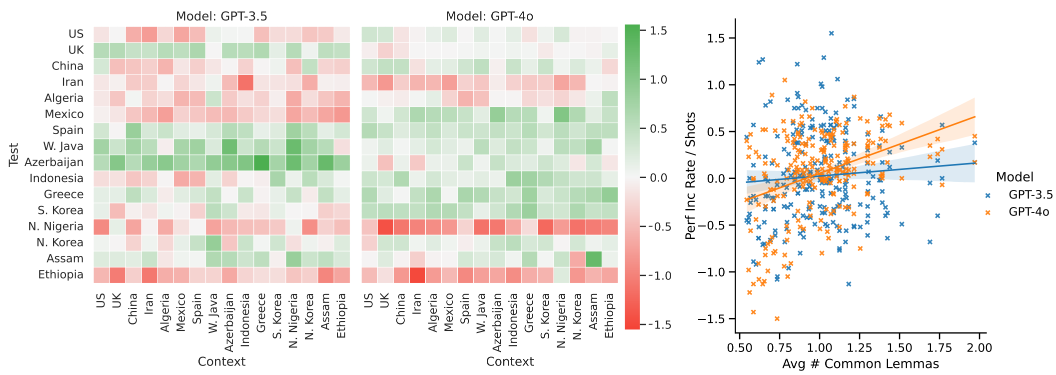

Figure 5: **Left:** Heat map of the performance increase rate by the number of demonstrations with GPT-4o and GPT-3.5 on BLEnD. **Right:** Correlation between performance increase rate and cultural proximity for each pair of cultures. Shaded bands indicate 95% confidence intervals.

5 Conclusion and Future Work

We investigated the effectiveness of in-context learning (ICL) in enhancing the cultural understanding of Large Language Models (LLMs). We demonstrated that culturally relevant in-context demonstrations generally improved the models' performance. We also observed that cross-cultural demonstrations could outperform same-culture demonstrations, indicating the potential for well-known cultural knowledge to support the understanding of less-represented ones. Our results highlight the importance of carefully selecting cultural demonstrations in building multi-cultural LLMs.

While we show the potential of ICL in enhancing cultural understanding within LLMs, we acknowledge certain limitations in our work. Our experiments were conducted using two closed-source LLMs, and further research is needed to validate whether the observed findings generalize to other models. Additionally, exploring parameter-efficient tuning on different cultural samples can provide valuable insights into cross-cultural generalization of LLMs, complementing or extending the capabilities demonstrated through ICL.

One promising future research is modeling relationships between cultures, enabling LLMs to leverage cross-cultural knowledge better. By understanding cultural proximity, overlap, or contrast, LLMs can more effectively incorporate demonstrations from well-known or neighboring cultures to improve performance on tasks associated with underrepresented cultures.

6 Social Impacts Statement

This research aims to enhance the LLM's understanding of diverse cultural knowledge. We believe that this kind of endeavor can contribute to more inclusive AI that respects the richness of global cultures. However, there are ethical considerations in ensuring that cultural demonstrations are carefully handled to avoid harmful stereotypes or misrepresentations. Additionally, the selective use of cross-cultural knowledge to improve performance in underrepresented cultures can lead to unintended consequences, such as reinforcing hierarchical views of cultural importance. It is crucial for future research to consider these ethical dimensions for responsibly incorporating cross-cultural knowledge into LLMs.

References

[1] Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural alignment of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.acl-long.671`.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,

Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

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

[4] Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, 2024.

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

[6] Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, 2022.

[7] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL `https://aclanthology.org/2022.emnlp-main.759`.

[8] Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. *arXiv preprint arXiv:2406.11661*, 2024.

[9] Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*, 2024.

[10] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

[11] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, 2022.

[12] Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*, 2024.

[13] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of large language models. *arXiv preprint arXiv*, 2311, 2024.

[14] Bin Wang, Geyu Lin, Zhengyuan Liu, Chengwei Wei, and Nancy Chen. CRAFT: Extracting and tuning cultural instructions from the wild. In Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Laura Cabello, Yong Cao, Ife Adebara, and Li Zhou, editors, *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 42–47, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.c3nlp-1.4`.

[15] Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.acl-long.345`.

[16] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, 2022.

[18] Wenjing Zhang, Siqi Xiao, Xuejiao Lei, Ning Wang, Huazheng Zhang, Meijuan An, Bikun Yang, Zhaoxiang Liu, Kai Wang, and Shiguo Lian. Methodology of adapting large english language models for specific cultural contexts. *arXiv preprint arXiv:2406.18192*, 2024.

A Prompt Examples

The Same Culture

```
What is the most popular wheat-based food item in the US?
A. bread
B. couscous
C. flour tortillas
D. jeon
Answer:  A. bread

...

What is the most popular fruit in the US? Do not include explanations.
A. apple
B. durian
C. mango
D. orange
Answer:
```

Different Cultures

```
What is the most popular wheat-based food item in West Java?
A. bread
B. couscous
C. jeon
D. steamed stuffed bun
Answer:  A. bread

...

What is the most popular fruit in the US? Do not include explanations.
A. apple
B. durian
C. mango
D. watermelon
Answer:
```

No Labels

```
What is the most popular wheat-based food item in the US?

...

What is the most popular fruit in the US? Do not include explanations.
A. apple
B. durian
C. mango
D. orange
Answer:
```

Wrong Labels

```
What is the most popular wheat-based food item in the US?
A. bread
B. couscous
C. flour tortillas
D. jeon
Answer:  B. couscous

...

What is the most popular fruit in the US? Do not include explanations.
A. apple
B. durian
C. mango
D. orange
Answer:
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See §2, §3, and §4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Although there is no independent limitation section, this paper discusses future directions (§5) as a preliminary work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There is no theoretical contribution.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The dataset, models, and experimental method used are described.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We plan to release the code when publishing at the archival venue.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The dataset, models, and experimental method used are described.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Confidence intervals are indicated.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The types of LLM APIs are described.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: There is no violation of the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See §6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The assets we used are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.