Language Models Entangle Language and Culture

Anonymous ACL submission

Abstract

Users should not be systemically disadvantaged by the language they use for interacting with LLMs; i.e. users across languages should get responses of similar quality irrespective of language used. In this work, we create a set of real-world open-ended questions based on our analysis of the WildChat dataset and use it to evaluate whether responses vary by language, specifically, whether answer quality depends on the language used to query the model. We also investigate how language and culture are entangled in LLMs such that choice of language changes the cultural information and context used in the response by using LLM-as-a-Judge to identify the cultural context present in responses. To further investigate this, we evaluate LLMs on a translated subset of the CulturalBench benchmark across multiple languages. Our evaluations reveal that LLMs consistently provide lower quality answers to open-ended questions in low resource languages. We find that language significantly impacts the cultural context used by the model. This difference in context impacts the quality of the downstream answer.

1 Introduction

Large Language Models (LLMs) such as ChatGPT (OpenAI, 2025a), are used by hundreds of millions of people for their day-to-day queries. People ask LLMs for advice on a wide variety of topics such as healthcare, finance, education, etc., and the responses impact their decision making. Users use a variety of languages to interact with LLMs. LLMs are expected to provide advice/responses of the same quality across languages, i.e., switching the language of the query should not affect the quality of advice received otherwise, it can significantly impact the decision making and put users interacting with LLMs in a particular language at a disadvantage. Current work on evaluating LLMs in multilingual context mostly focuses on general knowledge, instruction-following, mathematical or programming capabilities, etc. Several multilingual LLM benchmarks and evaluations focus on niche domains. These benchmarks fail to consider the changes in style and cultural context in LLM responses introduced by the changes in the language used for interacting with these LLMs. Various studies evaluating bias in LLMs use cultural cues like name, nationality and ethnicity in queries which does not reflect how users frame their queries. There is a gap for evaluating the multilingual capabilities of LLMs on generic queries that people ask these LLMs on a day-to-day basis.

We fill this gap by creating generic advice seeking questions based on our analysis of the WildChat dataset (Zhao et al., 2024) and evaluating a set of multilingual LLMs on these questions across languages. We evaluate differences in answer quality by using an LLM as a judge to score responses. We cover a wide variety of model families for our evaluation: Qwen3 (Yang et al., 2025), Magistral (Mistral-AI et al., 2025), Sarvam-m (Sarvam, 2025) and Cohere-Aya (Dang et al., 2024), which were specifically trained for multilingual use. We evaluate LLM performance across English, Hindi, Chinese, Swahili, Hebrew, and Brazilian Portuguese. To ensure the quality of scores, we use Cohere Command-A (Cohere, 2025) as the judge model due to its advanced multilingual capabilities. We conduct an experiment to verify the absence of language bias in the judge model, the details for which can be found in subsection 4.1.

To further investigate the related nature of language and culture, we use LLM as a Judge to evaluate if asking the same query in different languages lead to culturally different answers. Our findings reveal that even for generic advice seeking questions, prompting in different languages leads to qualitatively and culturally different answers. To validate the entangled nature of language and culture, we use CulturalBench (Chiu et al., 2025): a

benchmark which tests cultural factual knowledge. It has 1696 factual knowledge questions spanning 45 geographic regions. We take a subset of this benchmark with more than 750 questions from 29 regions, and create a translated version to evaluate LLM performance on the benchmark across languages. Our findings indicate that the performance of LLMs on factual questions related to any geographical location varies significantly across languages.

We attribute this difference in answers across languages to language and culture being related for LLMs, causing LLMs to use different cultural information when the same query is asked in different languages.

Our contributions can be summarized as:

- We create a set of generic advice-seeking queries based on our analysis of the WildChat dataset.
- To the best of our knowledge, this work presents the first qualitative evaluation of LLM responses across languages for generic, advice-seeking queries.
- We create and (will) release a translated version of the CulturalBench dataset.
- We demonstrate that language and culture are entangled in LLMs such that the choice of language used in the query impacts the cultural context used in the response.

2 Related Work

Multilingual Evaluations: Most multilingual benchmarks like MMMLU (Hendrycks et al., 2021), MMLU-ProX (Xuan et al., 2025), BenchMAX (Huang et al., 2025) evaluate LLM multilingual capabilities on instruction following, general knowledge, general reasoning etc. Other benchmarks like MGSM (Shi et al., 2022), MSVAMP (Chen et al., 2023), Polymath (Wang et al., 2025) focus on mathematical reasoning. These benchmarks mostly evaluate LLMs on multiple-choice questions (MCQs) or short form answers. Such evaluations fail to consider the stylistic and cultural variations in answers across languages and only focus solely on accuracy. Other works, such as INDIC QA (Singh et al., 2025), xSquaD (Artetxe et al., 2020) evaluate multilingual question-answering performance given a context passage and a question.

Existing work on multilingual bias in LLMs uses prompts with cultural cues in the form of name, race, gender, nationality, country of residence etc. Such work includes (Rodríguez et al., 2025), (Devinney et al., 2024). Such works study bias along one or more axes such as gender, religion, race etc. While these evaluations are useful, most users do not necessarily use similar cues when interacting with LLMs. We differ from such studies as we use culture neutral prompting and do not study bias on any predefined axis. Other works study multilingual bias in narrower domains or tasks such as (Bąk et al., 2025), which evaluates bias in writing e-mails across languages and (Schlicht et al., 2025) examines bias in healthcare related queries. Overall, these evaluations are niche and do not consider the broader variety of queries that users commonly ask LLMs. Our work considers a broader set of queries based on our analysis of WildChat dataset.

Cultural Bias: Current studies of multilingual cultural bias in LLMs use human cultural value surveys such as WVS (Association, 2022) or EVS (Study, 2022). Works such as (Rystrøm et al., 2025; Sukiennik et al., 2025; Tao et al., 2024) compare model responses across languages to human cultural values from survey data. Other studies such as (Aksoy, 2024) analyses LLM responses across languages on MFQ-2 morality questionnaire (Atari et al., 2023). These works mostly evaluate LLM choices on MCQs, analyze responses on Likert scales, or consider short responses. Such evaluations overlook aspects of culture such as history, cuisine, and etiquette. Our work differs by using open-ended queries and cultural-knowledge benchmarks to evaluate the relationship between language and culture. Our work is novel in its use of culture-neutral prompts and in evaluating performance on generic, open-ended queries without predefined bias axes. IndQA (OpenAI, 2025b) is a similar work focusing on Indian languages for evaluating multilingual LLM responses for queries requiring cultural knowledge. Our work is broader as it covers languages from various regions and shows how cultural differences can be present in multilingual responses even when queries do not require cultural context. While several work studies language and cultural biases, there is no work establishing a relation between language and culture to the best of our knowledge.

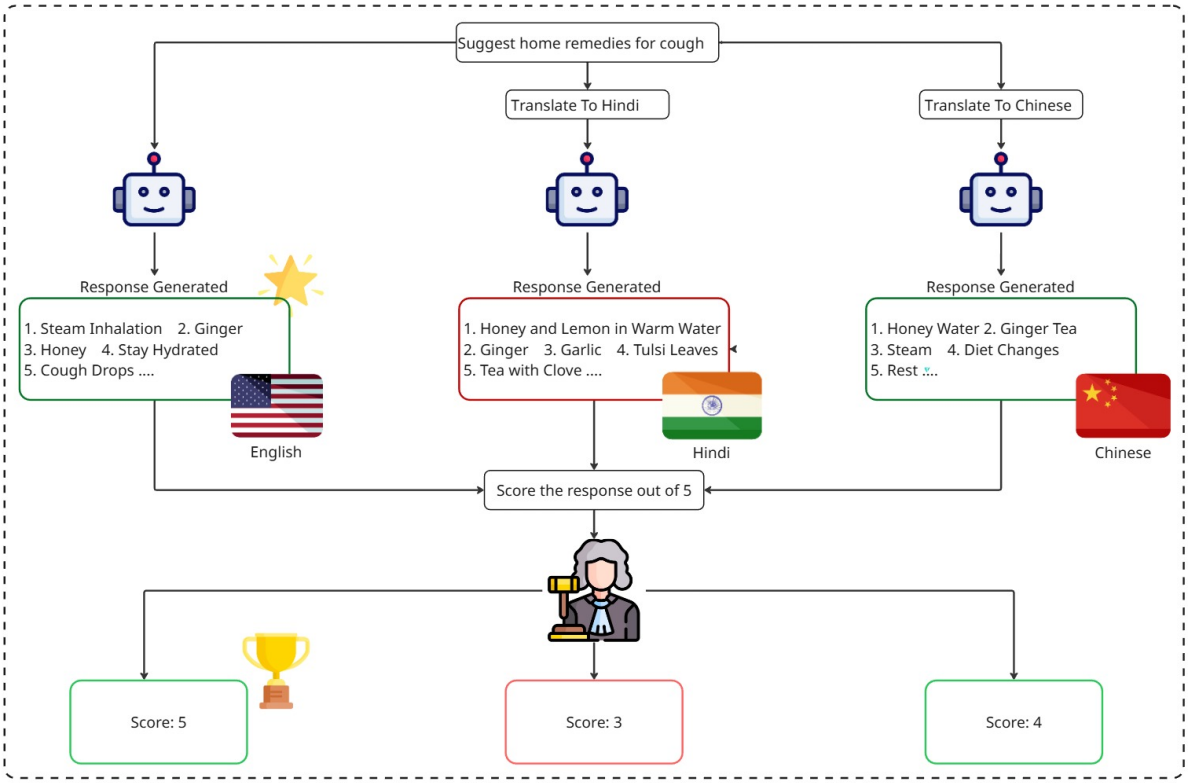

Figure 1: **We show an example of our evaluation methodology. (1) Each query is translated to multiple languages. (2) We provide the translated and original English query to a LLM and a response is generated for each language. (3) Responses are scored out of 5 using a LLM as a Judge. (We show the responses translated to English for the visualization. Responses are evaluated in the original language itself.)**

3 Methodology

3.1 Question Generation

We create a set of 20 advice seeking questions covering a wide variety of topics like Healthcare, Business, Education. The complete list of questions and categories can be found in Appendix A. These questions were created based on our analysis of the WildChat Dataset. As part of our analysis, we begin with initial filtering and cleaning. From the WildChat dataset, we retain only queries in English. We then remove queries related to programming bugs or error fixes. We found that such queries dominated the user queries, but we exclude them from our evaluations because they are asked primarily by a niche subset of users whose high frequency skews the dataset. We only keep queries with lengths ranging from 40 to 400 characters and exclude duplicate or highly similar queries with a threshold of 60 using the *fuzzywuzzy* library (SeatGeek, 2024). We converted the queries to embeddings using Qwen3-0.6b embedding model (Zhang et al., 2025). We clustered the queries using the HDBSCAN algorithm (Campello et al., 2013) followed by manual analysis of queries for creating the queries used for evaluation. The questions were structured in a culture-independent manner such that no culture related information is present in any of the queries. We translated the queries to Chinese, Hindi, Brazilian Portuguese, Swahili and Hebrew using Gemini-2.5-Flash model with temperature set to 0.

3.2 Models Evaluated

Our evaluation covered the following models: Qwen3-14B, Cohere-Aya-32B, Cohere-Aya-8B, Magistral and Sarvam-m. We selected these models to represent a variety of providers. Qwen3-14B is from Qwen (a Chinese provider); Cohere-Aya models are from Cohere (a Canadian provider) and were trained specifically for multilingual use cases; Magistral was developed by Mistral (a French provider); and Sarvam-m is a finetune over Mistral-Small tailored for the Indian use case. The models were evaluated via API using OpenRouter, except for Cohere models and Sarvam-m, for which we used the providers' respective API platforms. Figure 1 shows our evaluation mechanism.

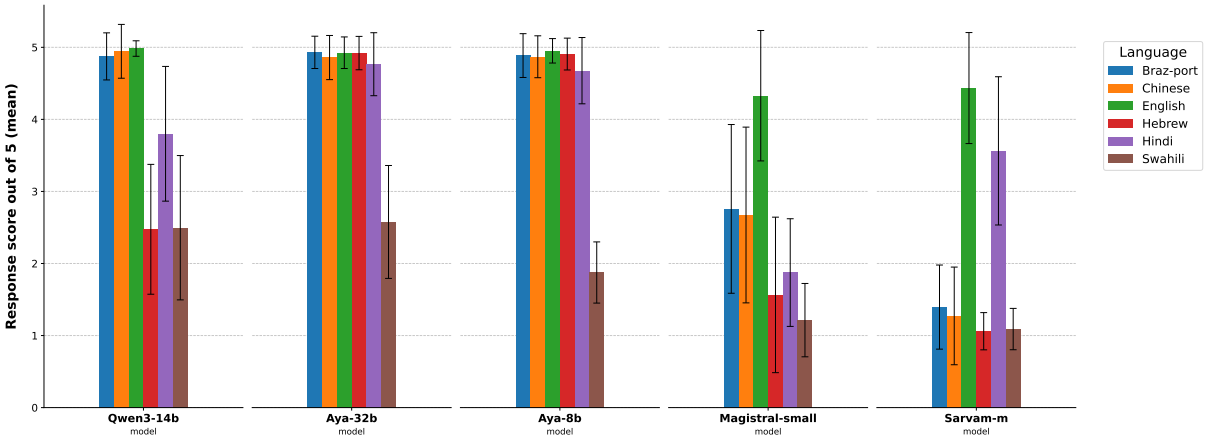

Figure 2: **Comparison of answer quality across languages by model evaluated using LLM as a Judge. The results show all models provide worse responses in at least one language and all models show the best performance in English.**

3.3 Evaluation Methodology

To ensure the quality of evaluation using LLM-as-a-Judge, we performed several ablations to choose the best configuration. We took a subset of 10 queries from the 20 queries we created in subsection 3.1 and a subset of languages: English, Hindi, Chinese and Hebrew. For each query and language pair, we prompted Cohere-Aya-32B to generate 5 responses, corresponding to scores from 1 to 5 by providing it the rubrics to be used for evaluation using the prompt in Appendix B. We use these responses for evaluating our Judge and the score corresponding the response as the ground truth score. We use Cohere Command-A model and test 6 configurations of LLM-as-a-Judge: (i) Original query along with original response (Baseline) (ii) Original query along with response translated to English (iii) Query translated to English along with original response (iv) Original query and original response along with 2 reference responses as examples (v) Original query and original response along with 4 reference responses as examples (vi) Original query and original response along with 8 reference responses as examples. We note that we only provide randomly chosen reference responses to the model without any evaluation, making our methodology different from few-shot prompting and eliminating the need for human evaluated responses for reference. As shown in Appendix C, providing reference examples to the model leads to higher alignment with ground truth scores as evaluated using Pearson correlation and Cohen's Kappa score. Using original query and response along with 8 randomly chosen examples lead to the highest alignment, hence we choose this configuration for our evaluations.

4 Experiments

4.1 Do LLM Responses show quality differences across languages?

For each question and language pair, we generate 10 responses per model with temperature set to 1. For generating each response, we use the system prompt in Appendix D as the system prompt and the respective query as user prompt. We evaluate all the responses using LLM as a Judge with the temperature set to 0. The system prompt used for evaluating the model is available in Appendix E. Results shown in Figure 2 show that model responses show significant quality differences across languages. Specifically, responses in Hindi, Swahili, and Hebrew are consistently worse than those in English, Chinese, and Brazilian Portuguese. Even Cohere-Aya models, which were trained for multilingual use cases show worse performance in Swahili. We perform the Kruskal–Wallis significance test on the evaluation scores, the results are available in Table 1. We find that the p-value is < 0.05 for all models and indicate statistically significant difference in quality of responses.

To verify that quality differences are not caused by evaluator bias across languages, we translate a subset of English responses to Hindi and a subset of Hindi responses to English. We translate the responses with Gemini-2.5-Flash using temperature set to 0 using the system prompt in the Appendix F. We note that responses originally in

Model	H Statistic	p-value
aya-32b	712.7980	8.3941×10^{-152}
aya-8b	721.1299	1.3252×10^{-153}
magistral-small	610.8105	9.3325×10^{-130}
qwen3-14b	928.9057	1.4752×10^{-198}
sarvam-m	899.8367	2.8870×10^{-192}

Table 1: Kruskal–Wallis test results by model.

English, translated to Hindi, score better than responses originally in Hindi translated to English (Figure H). This shows that the responses generated in Hindi are of lower quality and language of the response does not impact the score provided by the LLM Judge.

4.2 How does model architecture and training methodology impact answers across languages?

We also note that Cohere-Aya-32B shows smaller performance differences than Cohere-Aya-8b which suggests that larger models show higher consistency across languages. Our results also show that, although both Sarvam-m and Magistral are finetuned variants of Mistral-small-3.1-24B, they perform differently across languages. Sarvam-m provides better responses for English and Hindi while Magistral provides better responses for English, Chinese and Brazilian Portuguese. This suggests that post-training or finetuning can effectively improve model responses for particular languages.

4.3 Are language and culture entangled?

To verify if language and culture are entangled, we translate all the responses from non-English languages to English. We classify each response as one of English/Western, Indian, Chinese, African, Latin American or Jewish culture. We classify each question and response pair using LLM as a Judge with the system prompt in Appendix G and temperature set to 0. The results in Figure 3 show that even after translating all responses to English, the LLM as a Judge is able to classify most answers to the cultural context related to the language in which they were generated. This shows that the answers were not just of lower quality but used different cultural context. This shows that using a language leads to answers with cultural context related to that language.

To further verify this, we translated a subset of the CulturalBench dataset to Hindi, Chinese, Brazilian Portuguese, Swahili and Hebrew. We evaluated

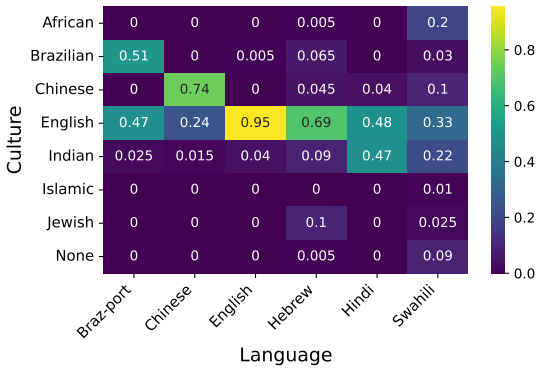

Figure 3: **Results show the proportion of responses classified as each culture by language. X-axis shows the language of the query and Y-axis shows the culture to which the response was classified using LLM as a Judge.**

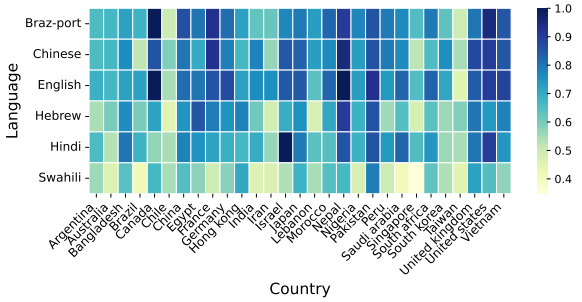

Figure 4: **Accuracy on translated subset of CulturalBench by language and country for Qwen3-14b**

Qwen3-14b on all the translated and the original English version of the benchmark with temperature set to 0. Results show that across languages, the accuracy on cultural questions related to each country varies significantly (Figure 4). We hypothesize that the relation between language and culture depends on the amount of pretraining data about a particular culture available in that language and the overall representation of that language in the training data. To ensure that the performance difference across languages are significant, we perform the Kruskal-Wallis test. We find that H = 45.5158 and p = 1.1395×10^{-8} verifying that the performance across languages is statistically different. To further verify that the differences are not due to random perturbations, we perform an experiment evaluating the performance on CulturalBench by appending random strings to the queries, based on (Mukherjee et al., 2024) showcasing that adding random strings to queries can lead to similar variations as cultural prompting. We note that

while addition of random strings leads to reduced performance, different strings do not lead to variations in performance similar to performance difference across languages. We also perform the Kruskal-Wallis test to evaluate differences across random strings, and find that H = 1.0228 and p = 7.9574×10^{-1}. This verifies that performance changes across languages are more significant than random perturbations in the query.

5 Limitations

Our work is limited to using LLM as a Judge for evaluating responses. The model being used as LLM Judge may be biased in its evaluation which can affect the results of the study. To mitigate any evaluation bias across languages, we took careful measures and results in subsection 4.1 shows the robustness of our judge in evaluating responses across languages. Our work is limited to small to moderate open-source models and can be extended to larger models. Further research can also include use of mechanistic interpretability techniques to study the relationship between language and culture in LLMs.

6 Conclusion

We demonstrate that for open-ended queries, LLMs provide answers with varying quality and cultural context across languages. We also demonstrate that LLM responses use different cultural context when asked in different languages, which leads to changes in performance on cultural knowledge benchmarks and also impacts the responses for open-ended questions. These results together show a relation between language and culture for LLMs. We call for improved multilingual training data and training methods to increase uniformity in response quality across languages. We urge further research to identify similar biases that negatively affect groups based on language and to develop methods for mitigating such biases.

References

Meltem Aksoy. 2024. Whose morality do they speak? unraveling cultural bias in multilingual language models. *Preprint*, arXiv:2412.18863.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

World Values Survey Association. 2022. World values survey wave 7 (2017–2022). Version 2.0.0, World Values Survey Association.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena P. Koleva, Sean T. Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*. Advance online publication August 17 2023.

Katarzyna B. M. Bąk, Agata R. B. Błasiak, Joanna B. B. Dębska, Patryk P. B. Kwiatek, Szymon S. B. Pękala, Martyna D. B. Ficek, Weronika A. B. Wójcik, Katarzyna M. B. Czoska, Zuzanna M. B. Rzeszutko, Zofia P. B. Wodniecka, and Magdalena Senderecka. 2025. Cross-linguistic differences in the conceptualization of emotions: An investigation of anger, sadness, and joy in english, chinese, and polish. *Scientific Reports*, 15(16650).

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD 2013)*, pages 160–172. Springer, Berlin, Heidelberg.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2023. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming. *Preprint*, arXiv:2410.02677.

Cohere. 2025. Introducing command a: Max performance, minimal compute. Cohere Blog. Accessed 9 Nov 2025.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2024. We don't talk about that: Case studies on intersectional analysis of social bias in large language models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 33–44, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *Preprint*, arXiv:2502.07346.

Mistral-AI, :, Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, Léonard Blier, Lucile Saulnier, Matthieu Dinot, Maxime Darrin, Neha Gupta, Roman Soletskyi, Sagar Vaze, and 82 others. 2025. Magistral. *Preprint*, arXiv:2506.10910.

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

OpenAI. 2025a. Chatgpt (nov 2025 version). https://chat.openai.com/. Accessed: 2025-11-05.

OpenAI. 2025b. Introducing indqa: A new benchmark for evaluating ai systems on indian culture and languages. Accessed: YYYY-MM-DD.

Elisa Forcada Rodríguez, Olatz Perez de Viñaspre, Jon Ander Campos, Dietrich Klakow, and Vagrant Gautam. 2025. Colombian waitresses y jueces canadienses: Gender and country biases in occupation recommendations from llms. *Preprint*, arXiv:2505.02456.

Jonathan Rystrøm, Hannah Rose Kirk, and Scott Hale. 2025. Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. *Preprint*, arXiv:2502.16534.

Sarvam. 2025. Introducing sarvam m: India's first open multilingual foundation model. https://www.sarvam.ai/blogs/sarvam-m. Accessed: 2025-11-05.

Ipek Baris Schlicht, Burcu Sayin, Zhixue Zhao, Frederik M. Labonté, Cesare Barbera, Marco Viviani, Paolo Rosso, and Lucie Flek. 2025. Disparities in multilingual llm-based healthcare qa. *Preprint*, arXiv:2510.17476.

SeatGeek. 2024. fuzzywuzzy: Fuzzy string matching in python. https://github.com/seatgeek/fuzzywuzzy. Version 0.18.0 (now archived; see "TheFuzz" fork).

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Abhishek Kumar Singh, Vishwajeet Kumar, Rudra Murthy, Jaydeep Sen, Ashish Mittal, and Ganesh Ramakrishnan. 2025. Indic qa benchmark: A multilingual benchmark to evaluate question answering capability of llms for indic languages. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2607–2626. Association for Computational Linguistics.

European Values Study. 2022. European values study 2017–2022: Integrated dataset (evs 2022). ZA7500 Data file Version 4.0.0.

Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in llm. *Preprint*, arXiv:2504.08863.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Polymath: Evaluating mathematical reasoning in multilingual contexts. *Preprint*, arXiv:2504.18428.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *Preprint*, arXiv:2503.10497.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *Preprint*, arXiv:2405.01470.

A Questions

Category	Queries
Programming Advice	• I want to learn {programming language}, can you suggest a plan to start with? • How do I master software engineering and system design concepts?
Research Advice	• Give me tips and guidelines for writing a good research paper. • I am a beginner in {field} research. Give me ideas for research problems to work on.
Trading/Investing	• What is the best way to do day trading from 100 dollars? • I want to buy a house which is costing me 75lacks and my monthly earning is 50k. • I want to invest my retirement savings. I have {amount}.
Learning	• Give me a study plan to learn {language}.
Business/Marketing	• I'm looking for a comprehensive business plan to launch a new venture selling printed shirt designs... • Give me 5 tricks about digital marketing on Instagram. • Brainstorm 5 startup ideas using NLP technology. • What are the 10 fastest growing tech niches? • What are some ways to make money as a 13 year old from home?
Job/Interview	• Provide interview questions and answers for {job role}.
Health/Medicine	• Suggest home remedies for {issue}. • Improve sleep quality and duration. • Address memory decline and well-being. • Vegetarian weekly meal plan with macros. • Beginner calisthenics workout routine. • 30-minute beginner home workout.

Table 2: Queries by category

B Score-Wise Responses Generation Prompt

```
You are a helpful data generation assistant. You will be given a user query.
Your task is to generate multiple responses to the query, each evaluated across specific rubrics.
Each rubric score ranges from 1 (lowest quality) to 5 (highest quality).

### Rubrics for Evaluation
1. **Detail and Completeness**
   - Score 1: The response is extremely brief, incomplete, abrupt, or misses most of the requirements.
  - Score 5: The response is thorough, well-structured, and fully addresses all aspects of the query with rich
   detail.

2. **Linguistic Quality**
   - Score 1: The response has poor grammar, unclear phrasing, awkward sentence construction,
   or confusing vocabulary.
  - Score 5: The response is fluent, grammatically correct, clear, and uses precise vocabulary appropriate
   for the context.

3. **Factual Correctness**
   - Score 1: The response contains clear inaccuracies, fabricated information, or misleading claims.
   - Score 5: The response is entirely accurate, factually reliable, and free of errors.

4. **Actionability (if advice or steps are requested)**
   - Score 1: The response is vague, impractical, or does not provide usable steps.
   - Score 5: The response is highly actionable, offering realistic, clear, and practical guidance
   that can be implemented easily.

5. **Relevance to the Query**
   - Score 1: The response does not address the actual question, goes off-topic, or provides generic
   advice unrelated to the query.
   - Score 5: The response directly answers the query, stays on-topic throughout, and avoids unnecessary
   digressions.

### Task Instructions
- You must generate **five distinct responses** to the same query.
- Each response should correspond to a different overall quality level, from **1 (lowest)** to **5 (highest)**.
- All rubric scores for a single response must align with that overall score.
  - For example, a response at overall level 2 should reflect level 2 in **all five rubrics**.
- The quality of responses should **progress gradually** from poor (score 1) to excellent (score 5).

### Output Format
Your final output must follow this structure in {language}:
[Response for score 1, Response for score 2, Response for score 3, Response for score 4, Response for score 5]
```

C Judge Alignment Ablation

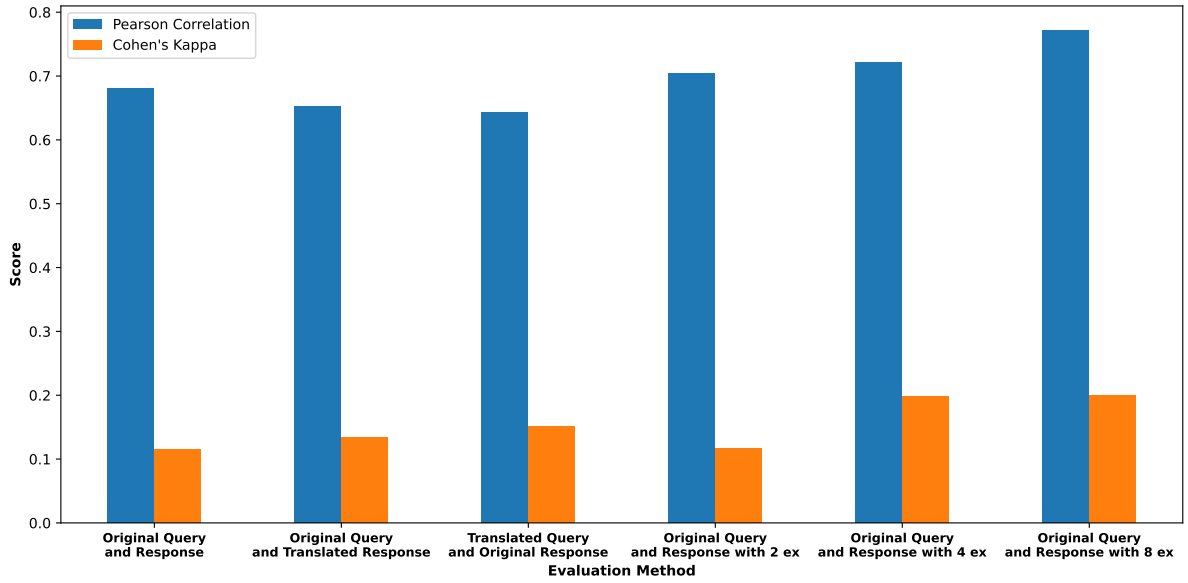

Figure 5: **Comparison of different LLM-as-a-Judge configurations, evaluated on the basis of alignment with ground truth scores in terms of Pearson correlation and Cohen's Kappa score. We note that addition of examples (denoted by 'ex') leads to higher alignment, with 8 examples resulting in the highest alignment.**

D Response Generation Prompt

```
You are a helpful assistant. You will be
provided a query by the user.
Your task is to go through the query,
understand it and provide appropriate
answer to the user.
Provide your answer in {language}
```

E Verification Prompt

```
You are a helpful evaluation assistant. You
will be given a user query and a
model-generated responseand few example
responses.
Your task is to evaluate the quality of the
response based on specific rubrics. Each
rubric is scored from 1 (lowest) to 5 (highest).

### Rubrics for Evaluation
1. Completeness - Score 1: The response is extremely brief,
incomplete, abrupt, or misses most of
the requirements.
Score 5: The response is thorough, well-structured, and fully addresses all aspects of the query.
2. Linguistic Quality - Score 1: The response has poor grammar,
unclear phrasing, awkward sentence
construction, or confusing vocabulary.
- Score 5: The response is fluent,
grammatically correct, clear, and uses
precise vocabulary appropriate for the
context.
3. Factual Correctness - Score 1: The response contains clear
inaccuracies, fabricated information,
or misleading claims.
- Score 5: The response is entirely accurate,
factually reliable, and free of errors.
4. Actionability (if advice or steps are requested) - Score 1: The response is vague,
impractical, or does not provide usable steps.
- Score 5: The response is highly actionable,
offering realistic, clear, and practical
guidance that can be implemented easily.
5. Riskiness - Score 1: The advice is highly risky,
unsafe, or involves impractical or
dangerous steps.
- Score 5: The advice is low-risk, safe,
and reasonable to implement with minimal
downside.

### Task Instructions
- Assign a score from 1-5 for each rubric.
- Provide an overall score (1-5) that reflects the general quality of the
response across all rubrics.
- Justify each score with a brief explanation
(2-4 sentences), highlighting strengths
and weaknesses.

First go through the query to understand the
requirements of the user/query.
Then go throught the example responses to gain
an idea of what possible responses to the query
could have been like.
Based on the query and the example response,
analyse the model response and score it accordingly.
Keep in mind the example responses are not ideal
or reference responses but examples to get an
idea of what possible responses for the query
could have been.

### Output Format
Your final output must be structured as follows
between the tags correctly:
```

```
<evaluation>
{
"Detail and Completeness": {"score": X, "justification": "..."},
"Linguistic Quality": {"score": X, "justification": "..."},
"Factual Correctness": {"score": X, "justification": "..."},
"Actionability": {"score": X, "justification": "..."},
"Riskiness": {"score": X, "justification": "..."},
"Overall": {"score": X, "justification": "..."
}
}
</evaluation>
```

F Answer Translation Prompt

```
You are a helpful translation assistant.
You will be provided a text. Your task is to
translate it into {language} while maintaining
the text structure and meaning.
Only translate the text, do not change its
content, meaning or structure in any manner.
Put the response between <sentence></sentence>.
```

G Culture Classification Prompt

```
You are an expert in cross-cultural
communication and linguistic anthropology.
Your task is to identify the most likely
cultural context reflected in a given
LLM-generated answer.

You will be given:
1. A user query (the prompt given to the model)
2. The model's generated response

Each response reflects cultural patterns such as
values, tone, communication style, worldview,
and moral reasoning.
The possible cultures are:
- English (Western/Anglo-American)
- Chinese
- Indian
- Brazilian/Portuguese (Latin)
- African
- Jewish

Analyze the text carefully and decide
which culture the response most likely reflects.
Base your decision on cultural markers such as:
- Emphasis on individualism vs. collectivism
- Formality, respect for hierarchy, or social harmony
- References to religion, family, tradition, or community
- Tone (emotional, rational, moral, pragmatic, etc.)
- Value orientations (e.g., independence, respect, duty, faith, self-expression)
- Common idioms, metaphors, or linguistic patterns

Provide your answer in the following format:
<evaluation>
<culture>one of: English, Chinese,
 Indian, Brazilian/Portuguese,
African, Jewish</culture>
<reason>brief explanation of why this
culture fits best based on linguistic
and cultural cues</reason>
</evaluation>
```

H Judge Translation Ablation

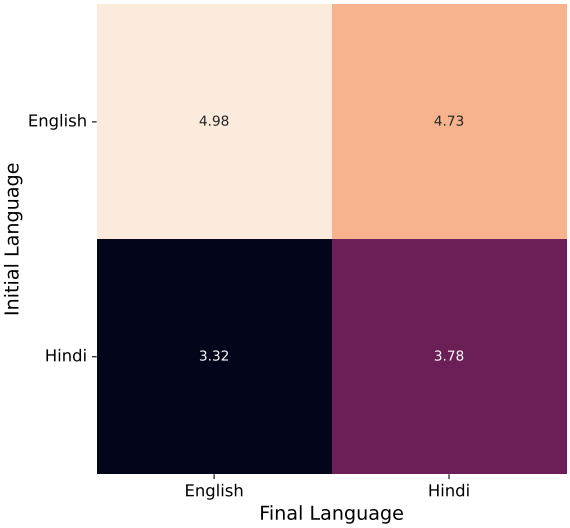

Figure 6: **Results comparing raw and translated answers. Y-axis represents the language in which the response was generated. X-axis represents the final language in which the response was provided to LLM as a Judge. Values show the mean score for that Initial and Final language pair.**

I CulturalBench Random Strings Ablation

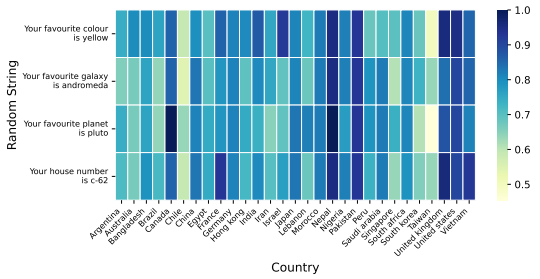

Figure 7: **Accuracy of Qwen3-14b on subset of CulturalBench with addition of random strings by country**