Toward a Culturally Competent LLM: Evaluating and Training Models for Global Understanding

Andrew Wesel[1] and Gheed El Bizri[1]
[1]Stanford University
Equal contributions. Correspondence: awesel@stanford.edu

Introduction

- Large language models (LLMs) have rapidly expanded in capability and now serve roles like therapists, employees, and legal assistants.
- Their widespread use means LLMs interact with a broad and diverse global audience.
- Despite this global reach, cutting-edge LLM development is concentrated in a few locations like the San Francisco Bay Area and Hangzhou, China.
- This geographic concentration raises concerns about whether models reflect the cultural diversity of the populations they serve.
- We aim to benchmark and improve cultural understanding in LLMs by turning it into a measurable task.
- Previous work (e.g. BLEnD) has focused on factual question answering, but we propose using survey re-creation as a way to evaluate cultural alignment.
- This project connects to broader efforts to evaluate subjective performance in AI, a field lagging behind math and coding tasks in quantifiability.
- By developing ways to give meaningful feedback on cultural sensitivity, we hope to close that gap and improve LLM performance globally.

Dataset

We utilize GlobalOpinionQA, a dataset from Anthropic that contains 6500 questions and national distributions of answers taken from the General Social Survey and other large cross-national surveys. Then, we queried different leading model APIs, asking the model to reason about a country's response distribution to a question, and predict what it might be. We leverage the reasoning capabilities of large language models, encouraging, through our prompt, the model to consider its knowledge about a national population using <think> tags, then return its final answer in brackets.

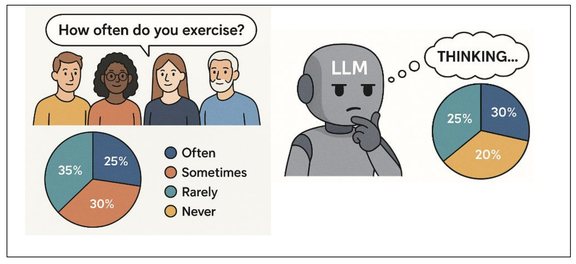

Fig.2: LLMs attempt to create national survey distributions through reasoning.

Our prompting approach differs in two ways from the way Anthropic used this dataset to elicit LLM global values.

1. Instead of asking the LLM the same question several times to give a single answer, and combine those answers to elicit a distribution, we ask the LLM once to give the entire distribution. We observe generally lower error with this method.
2. We ask the model to give reasoning and consider options and then answer, instead of answering, then justifying an answer. Language models use writing/token generation to think, and predicting the answer in the first token compresses all relevant computation into that first token instead of distributing it.

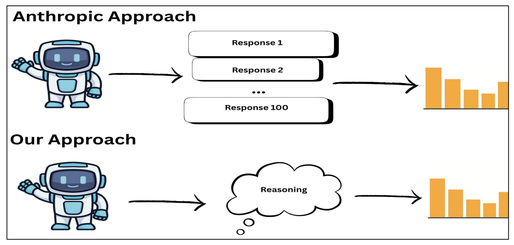

Fig.3: The differences between our processes.

Results

Benchmarking

- Used Jensen-Shannon distance (JSD) to measure divergence between each model's predicted answer distribution and true national survey distribution (range 0–1).
- Evaluated six models on 600 examples drawn from six culturally diverse countries.
- Models compared included GPT-4.1-mini, Gemini 2.0-Flash, Claude 3.5-Haiku,, Qwen-2.5..
- Key result: GPT-4.1-mini achieved the lowest average JSD (~0.25); competing models showed higher JSDs (≈0.3)

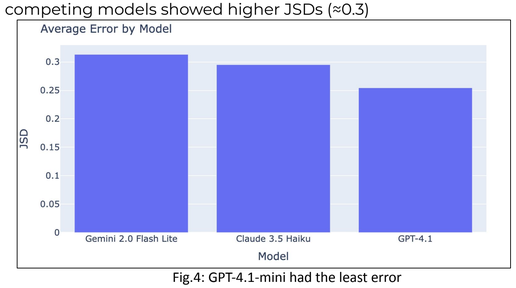

Fig.4: GPT-4.1-mini had the least error

Interpretability

- Ran GPT-4.1 on a larger 5800-question dataset (rather than the six-country, 600-question subset used for benchmarking).
- There are no obvious explanations for which countries the model performs well or poorly on. We examined language, internet usage, and cultural fractionalization but found no clear trends.

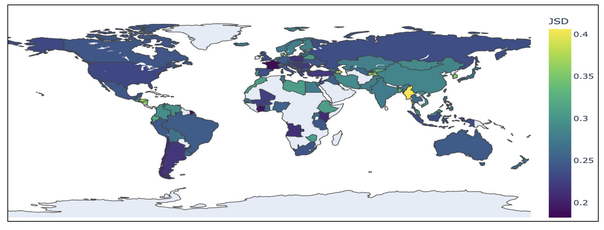

Fig.5: LLM error (JSD) on different countries. Gray: No Data

- This led us to investigate structural features of our questions.
- Intuitively, we can infer that some questions might be more difficult than others if they require more reasoning or specialized knowledge.
- We found that questions with more options (and therefore more opportunities for error) tend to have worse predictions.
- Since different countries have different average numbers of questions, this explains some of the divergence in performance.

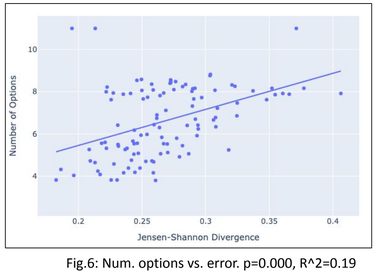

Fig.6: Num. options vs. error. p=0.000, R^2=0.19

Fine-tuning a custom model

- Goal: fine-tune a model that can outperform leading large models on this task
- Method: supervised fine-tuning using low-rank adaptation (LoRA)
- Benefit: LoRA allows a small number of data points to meaningfully change model performance
- Chosen base model: Qwen-2.5-7B-Instruct (top performer in benchmarking)
- Data selection: extracted GPT-4.1-mini's best reasoning traces and responses
- Low-error cutoff: JSD < 0.15
- Training dataset size: approximately 1,300 examples

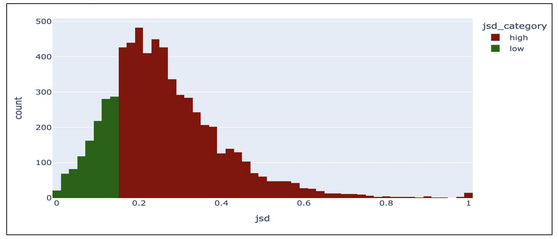

Fig.7: Our cutoff of low-error examples was at 0.15 JSD

- Rented a single NVIDIA H100 GPU from Lambda for training; total training time was about 30 minutes.
- Chose Qwen-2.5-7B-Instruct as the base model because its existing instruction tuning avoided re-learning answer formatting and kept focus on the task.
- Qwen-2.5 is among the strongest open-source models, especially for multilingual benchmarks.
- Used the same GPU for inference on roughly 600 queries (each containing hundreds of tokens), which took about four hours—illustrating that inference-time compute often exceeds training compute.
- When the fine-tuned model formatted its answer correctly, it significantly outperformed much larger leading models.
- However, the fine-tuned model failed to "box in" its answers 60% of the time, compared to only a 1% failure rate for the base Qwen-2.5-7B-Instruct model.

This finding highlights that fine-tuning on a small set of high-quality examples can introduce unintended downstream behavior.

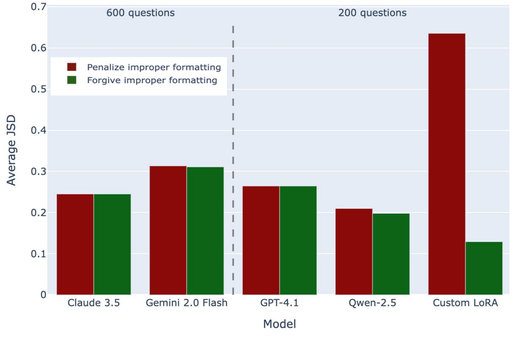

Fig.8: Our custom model outperformed every leading model, disregarding formatting errors. Gemini and Claude were evaluated on the 600-question benchmarking set. GPT, Qwen, and Custom were evaluated on the 200 questions not in training data.

Discussion

1] Benchmarking.

- In comparing each leading LLM provider's small model, we found that Qwen-2.5-7B had highest performance.
- Since this task inherently relies on reasoning (retrieval of information, weighting and logically combining facts retrieved into a properly-formatted final answer), we would love to continue this project by benchmarking models trained to reason, like OpenAI's o3 or Google's Pro series of Gemini models.
- Due to the excessive API costs (about 25x for larger models), we were limited to cheaper options and incentivized reasoning through chain-of-thought prompting, rather than reinforcement learning, as o3 or DeepSeek's R1 were trained.

2] Interpreting model performance.

- We found that, surprisingly, model performance did not correlate with "common sense" variables like internet penetration, language spoken, or ethnic fractionalization.
- To counteract this, we developed a "true score" that considers the number of options and weights harder questions more heavily.
- The map below shows the countries of the world scored with this new metric. Angola and Ivory Coast have much higher scores than any other country. We don't know why.

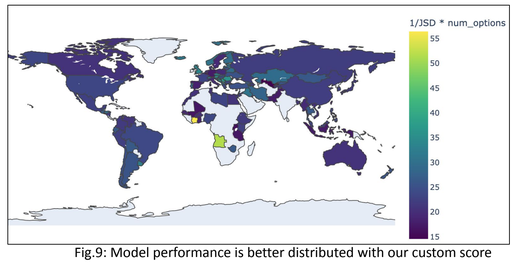

Fig.9: Model performance is better distributed with our custom score

3] Fine-tuning a custom model. Our model outperformed leading model providers on our own metric. Our training set represented 1200 high-quality responses, but our evaluation set was only 83 questions. This small size is due to multiple cuts; we removed all questions that were in the training set to prevent cross-contamination, which was 400/600, and we further removed questions that any model failed to format an answer properly which was 120/200. Needless to say, this is a small and highly conditional sample that does not perfectly represent performance on an abstract cultural reasoning task. We are curious about reinforcement learning as a way to teach our model to format its answers properly and potential improve reasoning as a result.

View outputs from our model

Purpose

Our study is a step towards developing models with a rich grasp of social contexts so they can serve and respect everyone, bringing us closer to truly culturally competent AI.

Select References

DeepSeek-AI. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv. https://arxiv.org/abs/2501.12948

Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., & Ganguli, D. (2024). *Towards measuring the representation of subjective global opinions in language models*. arXiv. https://arxiv.org/abs/2306.16388

Qwen Team. (2024, September). Qwen2.5: A Party of Foundation Models. Retrieved from https://qwenlm.github.io/blog/qwen2.5/