The MENAValues Shared Task on Cultural and Multilingual Alignment

Anonymous Author(s)
Affiliation
email@example.com

Abstract

This proposal outlines a shared task on evaluating and improving the cultural alignment of Large Language Models (LLMs) with the values of the Middle East and North Africa (MENA) region. The competition is based on the MENAValues Benchmark, a novel dataset derived from large-scale, authoritative human surveys. Participants will be challenged to develop models that not only accurately reflect the documented values of MENA populations but also maintain consistency across different languages and contextual framings. The task aims to foster innovation in creating more culturally aware and globally aligned AI systems, addressing a critical gap in current evaluation efforts. This proposal details the problem statement, the ethically sourced dataset, robust evaluation criteria, a strong baseline model, and a comprehensive plan for execution and publication.

1 Problem Statement

- **The ML Challenge:** Large Language Models often exhibit a Western-centric bias due to their training data, leading to significant cultural misalignment with non-Western populations. This shared task challenges participants to develop LLMs that are culturally aligned with the diverse values and beliefs of the Middle East and North Africa (MENA) region, which comprises over 500 million people. The core tasks are to (1) accurately predict population-level responses to value-based questions and (2) maintain response consistency across linguistic (English vs. Arabic, Persian, Turkish) and contextual (e.g., persona-based vs. observer) framings.
- **Impact:** As LLMs are deployed globally, their inability to reflect diverse cultural perspectives can erode user trust, reinforce stereotypes, and lead to harmful misrepresentations. Solving this challenge is crucial for building fairer, more inclusive, and trustworthy AI systems that can serve a global user base equitably.
- **Competition Format:** An ML competition is the ideal format to tackle this problem. It will galvanize the research community to develop novel techniques for cultural alignment, establish a robust, public benchmark for a critically underrepresented region, and encourage a diversity of solutions that move beyond monolithic, Western-centric models.

2 Dataset Considerations (Including Ethical Sourcing)

The competition will use the MENAValues Benchmark, a meticulously curated dataset designed for this challenge.

- **Source:** The dataset is built from two large-scale, high-quality, and publicly available survey datasets: the World Values Survey Wave 7 (WVS-7) and the Arab Opinion Index 2022

Preprint. Under review.

(AOI-2022). These are authoritative, professionally conducted surveys, ensuring ethical and legal compliance.

- **Representation:** The benchmark covers 16 MENA countries, including Egypt, Iran, Iraq, Jordan, Saudi Arabia, and Turkey. It contains 864 questions spanning four key dimensions: (1) Social & Cultural Identity, (2) Economic Dimensions, (3) Governance & Political Systems, and (4) Individual Wellbeing & Development.
- **Bias Mitigation:** The dataset is grounded in empirical human data, including full population-level response distributions. Rather than mitigating bias in the dataset, the goal is to use this "ground truth" data to identify and mitigate biases within the LLMs themselves. Post-stratification weights provided by the AOI-2022 survey are applied to ensure the data accurately reflects population demographics.
- **Privacy:** All data is derived from anonymized, publicly released survey results, ensuring no personally identifiable information is used.
- **Quality Assurance:**
  - **Human Sourcing:** The ground truth data is entirely human-generated through rigorous survey methodologies.
  - **Curation Quality:** Questions were manually selected for their relevance to cultural values. All benchmark questions were translated into Arabic, Persian, and Turkish and subsequently validated by native human annotators to ensure high quality and cultural nuance.
  - **GenAI Use:** No generative AI was used to create the ground-truth dataset. This is a benchmark of human values against which AI systems are to be measured.

3 Evaluation Criteria

The competition will feature two tracks with distinct evaluation criteria. Participants can submit to one or both.

- **Track 1: Cultural Value Alignment**
  - **Primary Metric: Normalized Value Alignment Score (NVAS)**. This measures the normalized absolute deviation between a model's predicted value and the ground-truth human average, with 100% indicating perfect alignment.
  - **Secondary Metric: Kullback-Leibler Divergence (KLD)**. This measures the dissimilarity between the model's full output probability distribution and the ground-truth human response distribution, rewarding models that capture the nuance of public opinion.
- **Track 2: Cross-Context Consistency**
  - **Primary Metrics: Framing Consistency Score (FCS)** and **Cross-Lingual Consistency Score (CLCS)**. FCS measures if a model's stance remains stable across persona-based and observer prompts. CLCS measures stability across English and native language prompts.
- **Justification:** This two-track structure allows for a comprehensive evaluation. NVAS and KLD measure a model's *authenticity* in reflecting cultural values. FCS and CLCS measure its *robustness* and cognitive coherence, directly addressing the key challenges of prompt-sensitive misalignment and cross-lingual value shifts identified in the original study.

4 Baseline and Current Performance

- **Baseline Model:** The baseline will be the Llama-3.1-8B-Instruct model, an open-source model evaluated in the original paper. It was chosen because the research found it offered the best balance of high value alignment (NVAS) and deep probabilistic alignment (KLD) among the open models tested.
- **Performance:** The zero-shot performance of the Llama-3.1-8B baseline is as follows:
  - **NVAS:** 75.75%

  - **KLD:** 1.31
  - **FCS:** 85.83%
  - **CLCS:** 79.30%
- **Purpose:** This strong, publicly available baseline provides a solid and reproducible entry point for all participants and clearly sets the performance bar to beat.

5 Platform

- **Preferred Hosting:** The competition will be hosted on CodaLab (`https://competitions.codalab.org/`), a widely-used platform for scientific machine learning competitions.
- **Rationale:** CodaLab provides excellent infrastructure for leaderboard management, automated evaluation, and handling private test sets, which is ideal for our two-track structure.
- **Commitment:**
  - The MENAValues dataset is permanently hosted on Hugging Face: `https://huggingface.co/datasets/llm-lab/MENA_VALUES_Benchmark` to ensure transparency and long-term availability.
  - All related code, evaluation scripts, and baseline implementations are available on the official GitHub repository: `https://github.com/llm-lab-org/MENA-Values-Benchmark-Evaluating-Cultural-Alignment-and-Multilingual-Bias-in-Large-Language-Models`.
- **Feasibility:** There are no known technical or legal constraints for using CodaLab with our publicly-grounded dataset.

6 Potential Positive Impact

- The competition directly supports the development of more fair, inclusive, and responsible AI by focusing on an underrepresented region.
- Anticipated benefits include the creation of more equitable and culturally aware NLP technologies, which can reduce the reinforcement of cultural stereotypes and improve user trust in global AI applications.
- The methodologies and models developed through this competition will provide a valuable template for evaluating and improving cultural alignment across other underrepresented regions, positively influencing the broader AI research community.

7 Proposed Competition Timeline (Post-Acceptance)

We commit that our dataset and codebase are ready, and that upon acceptance notification (mid-September 2025), the following fixed deadlines will be followed to deliver results and technical reports by mid-November 2025.

8 Plan for Publication

- **Post-Competition Paper:** We will publish a joint task overview paper summarizing the competition results, key findings, and analysis of top-performing participant approaches.
- **Co-authorship:** Top-performing teams will be invited as co-authors on the task overview paper.
- **Target Venues:** The initial results and system description papers will be published in the workshop proceedings. A follow-up, extended journal article will be considered for a relevant venue focusing on computational linguistics or AI ethics.

Date (2025)	Milestone	Details
Sept 18	Acceptance confirmed & announcement	Confirm scope, finalize rules; publish landing page and FAQs.
Sept 20	Release Train/Dev data & Baseline	Datasets on Hugging Face; code/scripts/baseline on GitHub; evaluation repo + starter kit released; platform page live with *public* dev leaderboard.
Oct 10	End of Development Phase	Public leaderboard frozen; prepare for test evaluation.
Oct 11	Release Test Set	Labels hidden; submission to *private* leaderboard begins.
Oct 18 (23:59 AoE)	Final Prediction Deadline	Private leaderboard closes; audit submissions for format, determinism, compliance.
Oct 24	Final Rankings Released	Standings published; notify teams; share CFP for system description papers.
Nov 3 (AoE)	System Description Papers Due	Teams submit 2 page technical reports (NeurIPS style).
Nov 10	Organizers complete editing	Reproducibility and ethics checks done; minor edits requested.
Nov 12	**Tech Reports Finalized**	Task overview report + accepted system descriptions finalized; PDFs camera-ready.
Dec 2	Workshop @ NeurIPS	Results session with overview talk, invited top teams, and panel.