Title

Cultural Persona Grounding and Evaluation: Enhancing Cultural Appropriateness in Large Language Model Personas

Problem Statement

Large language models often exhibit cultural biases and struggle to generate culturally appropriate personas and behaviors for diverse global contexts. This limitation hinders the development of inclusive and globally relevant AI characters, potentially leading to misunderstandings or offensive interactions in cross-cultural settings.

Motivation

Existing approaches typically focus on Western cultural norms or use simple token-based methods to adapt to different cultures, failing to capture deep cultural nuances. Our proposed Cultural Persona Grounding and Evaluation (CPGE) framework aims to address this gap by leveraging structured cultural knowledge and advanced machine learning techniques. By grounding personas in specific cultural contexts and evaluating their cultural appropriateness, we can create more inclusive and globally relevant AI characters that better represent and interact with diverse user populations.

Proposed Method

The CPGE framework consists of three main components: 1) Cultural Knowledge Base (CKB): A structured database containing information about cultural norms, values, and practices across different societies. 2) Grounding Module: A graph neural network-based system that takes a base persona description and a target culture, then generates a culturally-grounded version by reasoning over the CKB. 3) Evaluation Module: A multi-task learning system that assesses the cultural appropriateness of generated responses, simultaneously predicting cultural appropriateness scores and generating explanations for its judgments.

Step-by-Step Experiment Plan

Step 1: Data Collection and Preparation

Curate a diverse, multi-cultural dataset of persona descriptions and conversations, annotated by cultural experts. This dataset should cover at least 10 distinct cultures from different regions of the world. For each culture, collect at least 1000 persona descriptions and 5000 conversation snippets.

Step 2: Construct Cultural Knowledge Base

Build the CKB using existing cultural databases, academic literature, and expert knowledge. Structure the information as a graph, with nodes representing cultural concepts and edges representing relationships between them. Include at least 1000 nodes and 5000 edges per culture.

Step 3: Implement Grounding Module

Develop a graph neural network (GNN) model that takes a base persona description and a target culture as input, and outputs a culturally-grounded persona. Use the CKB as the underlying graph structure for the GNN. Train the model using the curated dataset, with the objective of maximizing the cultural appropriateness of the grounded personas.

Step 4: Implement Evaluation Module

Create a multi-task learning model that takes a persona description or a conversation snippet as input and outputs: a) A cultural appropriateness score (0-100), b) An explanation for the score. Train this model using the annotated dataset, optimizing for both score accuracy and explanation quality.

Step 5: Baseline Implementation

Implement baseline models for comparison: a) A simple token-based cultural adaptation method, b) A fine-tuned language model without explicit cultural knowledge. Use GPT-3.5 and GPT-4 as additional baselines.

Step 6: Evaluation

Assess CPGE and baselines on the following metrics: a) Cultural appropriateness (judged by native speakers), b) Ability to explain cultural differences, c) Performance on culture-specific tasks (e.g., navigating business etiquette in different countries). Use a held-out test set from the curated dataset for evaluation.

Step 7: Human Evaluation Study

Conduct a large-scale human evaluation study with participants from at least 20 different cultural backgrounds. Have them interact with personas generated by CPGE and baselines, rating their cultural authenticity and appropriateness.

Step 8: Analysis and Iteration

Analyze the results, identifying strengths and weaknesses of CPGE. Iterate on the framework, focusing on improving areas where it underperforms compared to baselines or human expectations.

Step 9: Ablation Studies

Conduct ablation studies to understand the contribution of each component (CKB, Grounding Module, Evaluation Module) to the overall performance of CPGE.

Step 10: Generalization Tests

Test CPGE on unseen cultures not included in the training data to assess its generalization capabilities.

Test Case Examples

Baseline Prompt Input

Generate a persona for a business professional in Japan.

Baseline Prompt Expected Output

Name: Hiroshi Tanaka Age: 42 Occupation: Senior Manager at a technology company Personality: Hardworking, punctual, and respectful of hierarchy Hobbies: Golf, reading business books, and

attending networking events Communication style: Formal and polite

Proposed Prompt Input

Generate a culturally grounded persona for a business professional in Japan, taking into account specific cultural norms and practices.

Proposed Prompt Expected Output

Name: Yuki Nakamura Age: 45 Occupation: Department Head at a traditional manufacturing company Personality: Diligent, group-oriented, and mindful of 'wa' (harmony) Hobbies: Participating in after-work nomikai (drinking sessions with colleagues), practicing calligraphy, and attending local matsuri (festivals) Communication style: Indirect and context-dependent, using keigo (honorific language) appropriately Cultural nuances: - Values long-term relationships and loyalty to the company - Practices nemawashi (consensus-building) before making decisions - Exchanges meishi (business cards) with proper etiquette - Respects senpai-kohai (senior-junior) relationships - Balances honne (true feelings) and tatemae (public facade) Typical workday: - Arrives early to prepare for the day - Participates in chorei (morning assembly) with team - Works long hours, often staying until the boss leaves - Engages in obligatory social activities with colleagues

- Values long-term relationships and loyalty to the company
- Practices nemawashi (consensus-building) before making decisions
- Exchanges meishi (business cards) with proper etiquette
- Respects senpai-kohai (senior-junior) relationships
- Balances honne (true feelings) and tatemae (public facade)
- Arrives early to prepare for the day
- Participates in chorei (morning assembly) with team
- Works long hours, often staying until the boss leaves
- Engages in obligatory social activities with colleagues

Explanation

The CPGE framework generates a more culturally nuanced and authentic persona compared to the baseline. It incorporates specific Japanese business practices (nemawashi, chorei), social concepts (wa, honne/tatemae), and cultural behaviors (nomikai participation, proper use of keigo) that are crucial for accurately representing a Japanese business professional. This level of cultural detail and authenticity is missing from the baseline output, which relies more on general stereotypes and lacks the depth of cultural understanding provided by the CPGE approach.

Fallback Plan

If the CPGE framework does not significantly outperform baselines, we will pivot to an in-depth analysis of cultural representation in language models. We will investigate where and why cultural misunderstandings occur, focusing on the following aspects: 1) Analyze the Cultural Knowledge Base to identify gaps or biases in our cultural data. 2) Examine the Grounding Module's performance across different cultures to understand if certain cultures are more challenging to represent accurately. 3) Investigate the Evaluation Module's criteria for cultural appropriateness and compare them with human judgments to identify discrepancies. 4) Conduct a comprehensive error analysis on cases where CPGE performs poorly, categorizing the types of cultural misrepresentations. This analysis could lead to valuable insights about the challenges of cross-cultural AI and inform future research directions in culturally-aware AI systems. Additionally, we could explore alternative approaches such as culture-specific fine-tuning of language models or developing a more sophisticated cultural reasoning system that goes beyond the current graph-based approach.