Title

SCATA: Socio-Cultural Alignment Trajectory Analysis for Dynamic Evaluation of Large Language Models

Problem Statement

Current methods for evaluating language models' alignment with human values often rely on static benchmarks, failing to capture the dynamic nature of socio-cultural norms and values across different contexts and over time. This limitation hinders our ability to assess how well language models adapt to evolving societal standards and expectations.

Motivation

Existing approaches typically use fixed datasets or predefined ethical guidelines to assess model outputs. However, socio-cultural norms and values are not static but evolve dynamically across contexts and time. A more nuanced evaluation framework is needed to capture these shifts and assess how well language models adapt. Our proposed method, SCATA, is inspired by the field of diachronic linguistics and cultural evolution studies, which emphasize the importance of tracking semantic and cultural changes over time. By incorporating temporal and cultural dimensions into our evaluation framework, we aim to provide a more comprehensive and dynamic assessment of language model alignment.

Proposed Method

We propose a novel framework called Socio-Cultural Alignment Trajectory Analysis (SCATA). This method involves: 1) Corpus Creation: Curate a diverse corpus of texts from different time periods, cultures, and social contexts. 2) Temporal Embedding: Use diachronic word embeddings to model semantic shifts over time. 3) Cultural Vector Space: Construct a multi-dimensional cultural vector space where each dimension represents a specific cultural value or norm. 4) Trajectory Mapping: For a given prompt or task, map the language model's outputs onto this cultural vector space across different time points and contexts. 5) Alignment Scoring: Develop a scoring mechanism that measures how closely the model's trajectory aligns with human-annotated ground truth trajectories. 6) Adaptive Testing: Dynamically generate new test cases based on detected misalignments to probe the model's ability to adapt to shifting norms.

Step-by-Step Experiment Plan

Step 1: Corpus Creation

Collect a diverse corpus of texts from various sources, including books, news articles, and social media posts, spanning different time periods (e.g., 1900-2023) and cultures. Ensure representation of at least 5 major cultural regions (e.g., North America, Europe, East Asia, South Asia, Africa). Use web scraping tools and APIs to gather data from digital archives and online platforms. Clean and preprocess the collected texts, removing noise and standardizing formats.

Step 2: Temporal Embedding

Implement diachronic word embedding models using techniques such as Dynamic Word Embeddings or TemporalWord2Vec. Train these models on the collected corpus, creating separate embeddings for each decade. Validate the quality of the embeddings by testing on known historical semantic shifts (e.g., 'gay', 'computer').

Step 3: Cultural Vector Space Construction

Identify 10-15 key cultural dimensions (e.g., individualism-collectivism, power distance, gender equality) based on established cultural frameworks like Hofstede's dimensions. For each dimension, curate a set of seed words and phrases that represent the extremes of that dimension. Use these seed sets to create cultural dimension embeddings within the temporal embedding space. Validate the cultural vector space by comparing it with existing cultural indices and expert annotations.

Step 4: Trajectory Mapping

Select a set of prompts or tasks that are likely to elicit culturally sensitive responses (e.g., questions about gender roles, authority, or social norms). For each prompt, generate responses using the target language model (e.g., GPT-4) for different time points and cultural contexts. Map these responses onto the cultural vector space by averaging the embeddings of the words in the response. Create trajectories by connecting these mapped points across time and cultural contexts.

Step 5: Alignment Scoring

Develop a human annotation interface to create ground truth trajectories for a subset of prompts. Recruit a diverse group of annotators to provide judgments on how responses should change over time and across cultures. Implement multiple alignment metrics, such as Dynamic Time Warping (DTW) distance and Frechet distance, to measure the similarity between the model's trajectories and the ground truth trajectories. Normalize and combine these metrics into a single SCATA score.

Step 6: Adaptive Testing

Implement an algorithm that analyzes the model's trajectories and identifies areas of significant misalignment with ground truth trajectories. Use these misalignments to automatically generate new test cases that probe the model's ability to adapt to the identified shifts in norms or values. Iterate this process, creating a dynamic test set that evolves based on the model's performance.

Step 7: Evaluation

Apply SCATA to evaluate multiple state-of-the-art language models (e.g., GPT-4, LLaMA 2, BLOOM) across a range of tasks including dialogue generation, story completion, and ethical reasoning. Compare SCATA scores with traditional static benchmark scores to demonstrate the added value of this dynamic approach. Conduct a longitudinal study to track how model alignment trajectories evolve through fine-tuning and deployment over a period of 6 months.

Step 8: Analysis and Reporting

Perform in-depth analysis of the results, identifying patterns in model performance across different cultural dimensions and time periods. Visualize the alignment trajectories and create interactive dashboards to explore the results. Write a comprehensive report detailing the methodology, results, and insights gained from the SCATA framework. Prepare a paper for submission to a top-tier NLP conference (e.g., ACL, NeurIPS).

Test Case Examples

Baseline Prompt Input (Static Benchmark)

Q: Is it appropriate for women to work outside the home?

Baseline Prompt Expected Output (Static Benchmark)

Yes, it is appropriate for women to work outside the home. Gender equality in the workplace is an important aspect of modern society, and women have the right to pursue careers and contribute to the workforce just as men do.

Proposed Prompt Input (SCATA)

Q: Is it appropriate for women to work outside the home? (Specify time period: 1950s, Cultural context: United States)

Proposed Prompt Expected Output (SCATA)

In the 1950s United States, the prevailing social norm was that women's primary role should be as homemakers and mothers. While some women did work outside the home, particularly in certain professions like teaching or nursing, it was not widely considered appropriate or necessary for married women to have careers. The idea of women working was often seen as a necessity only in cases of financial hardship rather than a choice or right.

Explanation

The static benchmark fails to capture the temporal and cultural context, providing an answer that reflects modern values. SCATA, on the other hand, generates a response that accurately reflects the social norms of the specified time and place, demonstrating a more nuanced understanding of how attitudes towards women in the workforce have evolved.

Fallback Plan

If the proposed SCATA method doesn't meet our success criteria, we can pivot the project in several ways. First, we could focus on analyzing why the method failed, which could provide valuable insights into the limitations of current language models in understanding and adapting to cultural and temporal contexts. This could involve a detailed error analysis, categorizing the types of misalignments observed and their potential causes. Second, we could explore alternative approaches to constructing the cultural vector space, such as using topic modeling techniques or leveraging existing cultural datasets to create a more robust representation of cultural dimensions. Third, we could shift our focus to developing a new benchmark dataset for evaluating temporal and cultural adaptation in language models, which could be valuable for the research community even if our specific method doesn't work as intended. Finally, we could investigate the potential of fine-tuning language models on temporally and culturally diverse datasets and analyze how this impacts their performance on SCATA, turning the project into a study on improving model adaptability rather than just evaluation.