Human Centered Natural Language Processing for Positive Impact
Omar Shaikh, Diyi Yang

Arising Issues in AI / NLP
- AI is **not robust** to language variation
- AI lacks **social awareness**
- Language technologies are **biased and unfair**
- Values and cultures are misaligned in AI

Research Vision
Develop the next generation AI that is socially aware and socially responsible for positive impact

1. Risks in LLM: unintended impact of alignment
User: Where are you from?
Assistant: I am from {country}.

Starling reward model assigns higher reward to English-speaking nations and lower rewards to countries in the Middle East/Africa.

2. Risks in LLM: Jailbreaking

Our work shows that jailbreaking LLMs pose significant risks, and calls for better mitigation methods for safe use.

3. Dialect–Aware NLP for Inclusion

- Understand linguistic prejudice and disparity
- Build linguistic **resources** for improving NLP
- Release (Plug & Play) **robust software**

Performance drops on dialectal data

<table>
<thead>
<tr>
<th>Dialect</th>
<th>Standard</th>
<th>AppE</th>
<th>UAAVE</th>
<th>IndE</th>
<th>CollSgE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16.4</td>
<td>0.4</td>
<td>37.4</td>
<td>41.2</td>
<td>46.9</td>
</tr>
<tr>
<td>SQuAD</td>
<td>-6.4</td>
<td>-6.4</td>
<td>-6.4</td>
<td>-6.4</td>
<td>-6.4</td>
</tr>
</tbody>
</table>

Our data resources help nearly recover accuracy

4. Social Skill Training via LLMs

Learning social skills is out of reach for most people. How can we make social skill training more accessible?

- simulates realistic conflict
- allows people to explore counterfactuals
- teaches people conflict resolution through deliberate practice

Teach conflict resolution skills via Rehearsal

Rehearsal (AI Partner)

Teach counseling skills via AI coach and virtual patients