1. Introduction: Normative Reasoning

- **Motivation**: Assistive and collaborative AI need to understand flexible (defeasible) social norms
  - largely missing from pre-training data
  - current models struggle to update inferences with new information beyond the most prototypical cases

- **Contributions**:
  - SCENE — a taxonomy for constraining norms within their sociocultural frame
  - NormBank — a knowledge bank of 155k situational norms
  - neural knowledge-completion models
  - empirical evidence for the utility of NormBank on downstream tasks

2. SCENE: A Dramaturgical Taxonomy

   ![Diagram of SCENE taxonomy]

   - **Setting**: restaurant, night, not crowded
   - **Environment**: literature review, prompting + filtering
   - **Roles**: customer, server
   - **Attributes**: age, role
   - **Behaviors**: drinking alcohol, going on a date

   - Knowledge Graphs: 129 setting constraints, 404 constraints, 928 role constraints, 578 attributes, 6,938 behaviors

3. Annotation: Building NormBank

   - **HIT**: start w/ behaviors and ask for contexts under which the behavior is:
     - highly creative task → norm: expected
     - scaffolded by SCENE → norm: unexpected

   - **Quality Control**:
     - **qualification test**: six questions on taxonomy definitions
     - **staging round**: small-scale task (not feeding into the final dataset) where workers get personalized feedback
     - **programmatic filtering, random audits**

4. Summary Statistics

   - **Distinct Constraints**: 408,407
   - **Taxonomic Constraints**: 93.6%
   - **Pre-Pop Constraints**: 63.2%
   - **Distinct Norms**: 155,423
   - **Social Situations**: 70,215
   - **Distinct Behaviors**: 6,938

5. Experiments: Automatic Expansion

   - **Classification**: given an unseen behavior + constraint set, predict the norm (expected, okay, unexpected)

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa</td>
<td>73.3%</td>
<td>71.4%</td>
<td>72.1%</td>
<td>75.4%</td>
</tr>
</tbody>
</table>

   - **Generation**: given an unseen behavior + norm label, generate constraints that make the norm label true

<table>
<thead>
<tr>
<th>Model</th>
<th>Sensible Norm</th>
<th>Correct Norm</th>
<th>Normative Constraint</th>
<th>Helpful Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-3</td>
<td>95.0%</td>
<td>61.1%</td>
<td>91.8%</td>
<td>87.8%</td>
</tr>
<tr>
<td>BART</td>
<td>100.0%</td>
<td>46.0%</td>
<td>94.3%</td>
<td>94.3%</td>
</tr>
<tr>
<td>Human</td>
<td>82.5%</td>
<td>55.0%</td>
<td>72.9%</td>
<td>81.7%</td>
</tr>
<tr>
<td>Max</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

6. Experiments: Downstream Transfer

   (via sequential fine-tuning)

7. Limitations + Ethical Considerations

   - **NormBank** is descriptive rather than prescriptive (NOT to give you advice)
   - **Annotator Demographics**: US English-Speakers
   - Future work can evaluate transfer utility on additional downstream tasks
   - **SCENE**’s role and attribute constraints will allow more targeted bias mitigation efforts