Problem Statement
- Question Answering (QA) is a common downstream task for NLP systems to test their ability to perform reading comprehension.
- The QA model takes as input a question and context and predicts the start and end positions of the answer.

Our Approach:
- **Data augmentation:** Generate new data for the out-of-domain dataset by backtranslating the context, question, and answer.
- **Masked Language Modeling:** Pre-train language models on question contexts to adapt to distribution shifts.

Background
- **Problem Setup:** We have input \((c, q)\) where \(c\) is the context and \(q\) is the query.
- Our goal is to predict the start and end indices \(s_{\text{out}}\) and \(e_{\text{out}}\) of the context which contain the answer \(a\) to the question.

DistilBERT:
- DistilBERT is a small-sized BERT model based on the Transformer architecture and uses a linear classifier for the QA head.
- The baseline is a pre-trained DistilBERT fine-tuned on training data.

Transfer Learning:
- Transfer learning helps address the distribution shift between the in-domain and out-of-domain datasets.
- We perform transfer learning by training on in-domain data and fine-tuning on out-of-domain data.

Backtranslation:
- We performed backtranslation on the out-of-domain dataset to generate more (context, query, answer) while preserving meaning.
- The best performing backtranslations were English → Hindi → English and English → Turkish → Hindi → English.

Masked Language Modeling:
- Perform pre-training with the MLM task, which requires the model to predict a randomly masked token in the input sentence.

Mixture of Experts:
- Train \(k\) experts and a gating function parameterized by an MLP.
- Each expert is trained on different subsets of the in-domain data.
- The gating MLP takes the embeddings produced by each expert and the tokenized data and outputs a weighted embedding.

Other Methods:
- Freezing: we experimented with freezing different layers of the network during fine-tuning on out-of-domain data.
- Regularization: we added a regularization term to the overall loss.

<table>
<thead>
<tr>
<th>Model</th>
<th>Race</th>
<th>RelEx</th>
<th>DuoRC</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>DistilBERT + FT</td>
<td>33.17(23.44)</td>
<td>64.13(45.31)</td>
<td>47.17(33.33)</td>
<td>48.90(34.02)</td>
</tr>
<tr>
<td>DistilBERT + FT + ME</td>
<td>31.46(20.97)</td>
<td>63.08(42.99)</td>
<td>46.19(27.06)</td>
<td>46.91(30.34)</td>
</tr>
<tr>
<td>DistilBERT + FT + BCT</td>
<td>38.81(25.78)</td>
<td>75.81(57.03)</td>
<td>42.81(34.13)</td>
<td>52.21(38.98)</td>
</tr>
<tr>
<td>DistilBERT + BCT + MLM</td>
<td>39.88(24.22)</td>
<td>76.23(60.16)</td>
<td>42.95(34.21)</td>
<td>53.02(39.53)</td>
</tr>
</tbody>
</table>

**Results:**
- **Test Dataset:** Final EM: 46.766 Final F1: 62.977 Rank: #3 [March 13]
- **Analysis:**
  - Relation Extraction particularly benefits by data augmentation.
  - Mixture of experts reduced scores across the board.
  - Backtranslation on languages with lower BLEU introduces variance.
  - MLM further helps the model adapt to the domain shift.

**Conclusions:**
- Customizing model training strategies and structures to out of distribution datasets significantly improves results and was the key to our success.

**References:**
1. [v2x_v2.x](http://example.com)
2. [v2x_v2.x](http://example.com)
3. [v2x_v2.x](http://example.com)
4. [v2x_v2.x](http://example.com)