**Problem Statement**

We work on building a **Robust Question Answering** system that can generalize to out-of-domain datasets with a small number of examples.

**Input:** Paragraph. **Question:** What is Tesla's stock price? **Context:** On 24 March 1879, Tesla was returned to Bosnia under police guard for not having a residence permit. On 17 April 1879, Mr. Tesla died at the age of 60 after contracting an unspecified illness (although some sources say that he died of a stroke).

**Answer:** not having a residence permit

---

**Method**

We use Model Agnostic Meta Learning (MAML) [1], which is an algorithm that trains a model to "learn how to learn".

We learn an effective representation of parameters $\theta$ that performs well on new tasks given few-shot training.

---

**Experiments**

We evaluate our models based on their EM and F1 scores, which are defined as follows:
- **EM Score**: a binary measure of whether the answer is correct = intuition: Is this exactly the actual answer?
- **F1 Score**: defined as $2 \times $precision $\times $recall / ($precision + $recall) = intuition: How close is the actual answer?

We demonstrate our methods on 4 candidate baseline models, with descriptions as follows:

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Model Description</th>
<th>Year-specific</th>
<th>Task-specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>DistilBERT baseline, no fine-tuning</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M2</td>
<td>DistilBERT baseline</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>M3</td>
<td>First-order MAML, on $M_0$</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>M4</td>
<td>Second-order MAML, on $M_0$</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

We notice that different datasets require different learning rates due to diverse underlying data characteristics, as demonstrated in Figure (6).

---

**Results & Analysis**

Based on our preliminary experiments, we had the following findings and followed up on them:

- We noticed that fine-tuning results were unstable, and therefore we tuned learning rates separately for each dataset, which helped increase stability significantly, as observed in Figures (8a, 8b).
- We noted significant variance in predictions across multiple seeds for the same model (Figure (9)), so we ensembled across seeds. This was observed to boost F1 scores, as seen in Figure (10).
- We performed data augmentation [3] to alleviate the lack of training data on out-of-domain sets; the advantage can be seen in Figures (7a, 7b).
- After looking at initial instability of meta-learned models, we proposed M4, where we used 2nd order MAML on top of a good pre-trained model based on ideas from the paper "How to train your MAML" [2].

---

**Future Work**

Due to high training costs, we were unable to find good hyperparameters for EDA and SO-MAML on training. We also expect that the following methods will help performance:

- Using OOD train data in the training step
- Incorporating importance sampling, even in just fine-tuning

---

**References**

2. "How to train your MAML". Jure Jakob Bresters, Hartoon Edward, and Arno Smeulier
3. "EDA: Easy Data augmentation techniques for boosting performance on text classification tasks", Levan Wae and Hie Zuo