**Research Architecture**

- **Pre-trained Language Models:**
  - RoBERTa-base
  - RoBERTa-large
  - BioBERT-base

- **Large Language Models:**
  - GPT-3.5
  - GPT-4.0

- **Fine-tuning Techniques:**

- **Evaluation Metrics:**
  - Accuracy
  - Precision
  - Recall
  - F1 Score

- **Best Performance Models:**
  - BioBERT-base
  - RoBERTa-large
  - GPT-3.5

- **1-shot Prompt Engineer**

**Prompt Engineer**

- **One-shot Learning:**
  - Analysis showed reduced accuracy in insight extraction from short, ambiguous tweets.
  - GPT models often over-labeled: GPT-3.5 labeled 929 instances, GPT-4.0 labeled 789, while the actual ground truth was 400.

**Fine-tuning Techniques**

- **Hyperparameters:**
  - Fine-tuned for multi-label text classification
  - Max Epochs: 10
  - Max Sequence Length: 512
  - Learning Rate: 5e-5
  - Batch Size: 16
  - Loss Func.: BCEWithLogitsLoss

**Prompt**

You are a medical expert analyzing tweets to check whether the user suffers adverse drug events.

**Scenario:**

- Because every text is from Twitter, the texts are short. Please consider this situation and annotate the text with proper labels to check whether the user suffers adverse drug events. For instance, users list the adverse drug effects rather than express personal experiences of adverse reactions.

**Your annotating steps are as follows:**

1. Check whether the user lists the adverse drug effects rather than expressing personal experiences of adverse reactions.
2. Check whether this tweet’s user suffers from adverse drug events.
3. Check the symptoms in these 22 symptoms listed below.

**Your annotation should be in the following format:**

1. If the user suffers from the tweet’s symptom instead of listing the adverse drug events, output with the corresponding label.
2. If the user doesn’t suffer from the symptom in the tweet, output with “None”.

**Symptom Labels:**
- nausea, diarrhea, ... rash, stomatitis

**Here is some annotate example for you to base on.**

*Text: I finished chemotherapy and side effects of the contrast dye. I was feeling kind of sick and nausea was getting worse, so I thought it would be tough, but this morning my chest hurts. I'm going to go to the hospital tomorrow, though I'm anxious because I have 2 hours left until the test results...*

*Label: nausea, pain*

**Other 22 samples for GPT to know.**

**Performance**

**Exact Accuracy of Test Dataset and Development Dataset**

<table>
<thead>
<tr>
<th>Models</th>
<th>Development Dataset</th>
<th>Test Dataset</th>
<th>Precision</th>
<th>ADE</th>
<th>NO ADE</th>
<th>Recall</th>
<th>NO ADE</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioBERT-Base</td>
<td>0.92</td>
<td>0.82</td>
<td>0.78</td>
<td>0.89</td>
<td>0.91</td>
<td>0.99</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Roberta-Base</td>
<td>0.76</td>
<td>-</td>
<td>0.72</td>
<td>0.70</td>
<td>0.71</td>
<td>0.99</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>RoBERTa-Large</td>
<td>0.85</td>
<td>0.81</td>
<td>0.83</td>
<td>0.67</td>
<td>0.71</td>
<td>0.99</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>GPT3.5</td>
<td>0.72</td>
<td>0.69</td>
<td>0.70</td>
<td>0.62</td>
<td>0.97</td>
<td>1.00</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>GPT4.0</td>
<td>0.63</td>
<td>-</td>
<td>0.57</td>
<td>0.60</td>
<td>0.99</td>
<td>1.00</td>
<td>0.98</td>
<td></td>
</tr>
</tbody>
</table>

**Subtask 1-5M-DAE-EN Binary and Per Label Performance Metrics**

<table>
<thead>
<tr>
<th>Models</th>
<th>Precision</th>
<th>ADE</th>
<th>NO ADE</th>
<th>Recall</th>
<th>NO ADE</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioBERT-Base</td>
<td>0.47</td>
<td>0.94</td>
<td>0.53</td>
<td>0.99</td>
<td>0.92</td>
<td>0.97</td>
</tr>
<tr>
<td>Roberta-Base</td>
<td>0.47</td>
<td>0.94</td>
<td>0.53</td>
<td>0.99</td>
<td>0.92</td>
<td>0.97</td>
</tr>
<tr>
<td>RoBERTa-Large</td>
<td>0.47</td>
<td>0.94</td>
<td>0.53</td>
<td>0.99</td>
<td>0.92</td>
<td>0.97</td>
</tr>
<tr>
<td>GPT3.5</td>
<td>0.47</td>
<td>0.94</td>
<td>0.53</td>
<td>0.99</td>
<td>0.92</td>
<td>0.97</td>
</tr>
<tr>
<td>GPT4.0</td>
<td>0.47</td>
<td>0.94</td>
<td>0.53</td>
<td>0.99</td>
<td>0.92</td>
<td>0.97</td>
</tr>
</tbody>
</table>

**Conclusions and Contributions**

- Expanded dataset with GPT-3.5 boosts RoBERTa accuracy from 0.76 to 0.86.
- Refined prompts for tweets, increasing GPT4.0 accuracy to 0.70.
- BioBERT excels in drug event extraction; GPT one-shot learning shows limits.

**Acknowledgment**

This research was supported in part by the National Science and Technology Council (NSTC), Taiwan, under grants MOST 110-2410-H-305-013-MY2, NSTC 112-2452-H-305-002., and NSTC 112-2627-M-038-001., and National Taipei University (NTPU). Taiwan under grants 112-NTPU-DKDA-F-001, 112-NTPU-DKDA F-004, 112-NTPU-DKDA F-001, 112-NTPU-DKDA F-005, and 112-NTPU-DKDA F-006.