Main Findings and Limitations

System A:

Validation data:

Epoch	Training Loss	Validation Loss

[1026/1026 02:30]

[8205/8205 55:35, Epoch 1/1]

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accuracy
1	0.025700	0.040220	0.905759	0.913828	0.909776	0.985417

Test data:

[1029/1029 02:08]

```
{'eval_loss': 0.027162281796336174,
 'eval_precision': 0.9366507177033493,
 'eval_recall': 0.954182101774225,
 'eval_f1': 0.9453351361792545,
 'eval_accuracy': 0.9904877170232692,
 'eval_runtime': 147.9741,
 'eval_samples_per_second': 222.39,
 'eval_steps_per_second': 6.954,
 'epoch': 1.0}
```

System B:

Validation data:

[8205/8205 55:41, Epoch 1/1]

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accuracy
1	0.012200	0.018261	0.951384	0.959614	0.955481	0.993665

Test data:

[1029/1029 02:07]

```
{'eval_loss': 0.016513075679540634,
 'eval_precision': 0.9616130283055447,
 'eval_recall': 0.9720830974260702,
 'eval_f1': 0.9668197175777528,
 'eval_accuracy': 0.9942540162812478,
 'eval_runtime': 145.2432,
 'eval_samples_per_second': 226.572,
 'eval_steps_per_second': 7.085,
 'epoch': 1.0}
```

Let's evaluate the fine-tuned models for System A and System B on their corresponding test data. It can be seen that even with one epoch, the models achieved an F1 score above 0.90 on validation data and even more on System B with a score of 0.95. Also, the models performed slightly better on their test data. The better performance of System B might be due to having fewer entities to classify

compared to System A. Further performance evaluation can be done with a confusion matrix to show models' entity-based performances. Also, another LLM can be fine-tuned.

Regarding limitations, time and computational resources are the main factors. Although the training process in this project covers only fine-tuning part of the pre-trained model, it takes time to train it for each epoch. Considering that my local GPU has limited memory, I used Google Colab resources. During the fine-tuning process, the training was stopped due to the limited time provided per user for GPU, leading me to train from the beginning with fewer epochs. With enough resources and time, comprehensive hyperparameter tuning can be achieved with a better validation approach, such as k-fold cross-validation, which might lead to more generalized models.