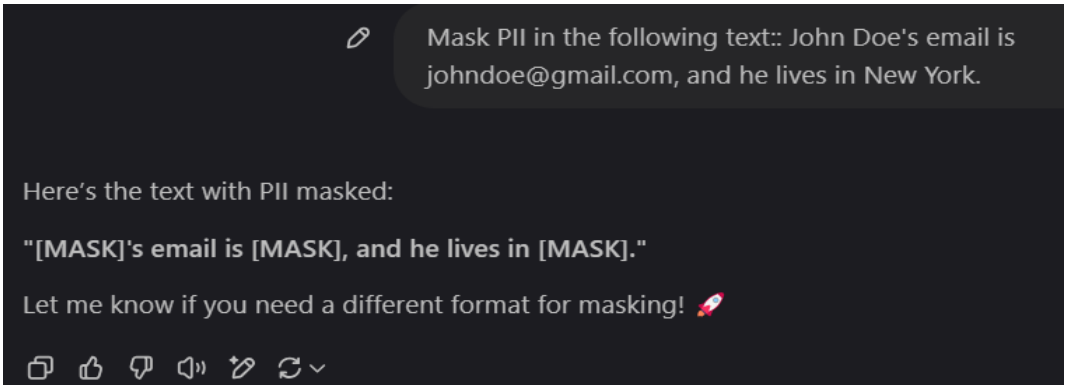
INTERNSHIP TASK REPORT

- The date was in English language only, and clean without any missing values or discrepancies.
- Emails were not present in the data. Thus, they were added to 50% of the dataset in somewhat realistic but random order.
    - Emails could be added realistically based on context via using some LLM prompting. But I chose to add them randomly because usually for NER data email is not tagged due to it being easily recognized by a Regular Expression, which is better suited to find them in the data rather than tuning an encoder.
    - Emails were only tagged as [B-EMAIL] completely to all tokens.
- Model was fine tuned on 4 tags ["O", "B-PER", "I-PER", "B-EMAIL"].
- Test_data was kept separate only for sole purpose of validation.
- "prajjwal1/bert-tiny" was used, a small and effective BERT variant freely available.

Epoch	Training Loss	Validation Loss
1	0.115800	0.075507
2	0.088000	0.060834
3	0.080100	0.058456

- Decreasing training and validation loss shows accurated fine tuning.
- Then, after evaluating on test_data, following metrics were computed:
    - **Precision (0.89)** – Good, indicates that most identified PII are correct with few false positives.
    - **Recall (0.91)** – Very good, suggests the model is capturing most PII instances with few misses.
    - **F1-score (0.90)** – Strong, balances precision and recall effectively.
    - **False Positive Rate (FPR - 0.0098)** – Excellent, very few non-PII are mistakenly masked.
    - **False Negative Rate (FNR - 0.083)** – Acceptable but reducing this further would ensure fewer PII are missed.
- Model was saved and a pipeline was created (Ref: “4- masking pipeline.ipynb”), tags were correctly identified by the model and the sentence was masked as per those tags.



(It is a prompt and answer from ChatGPT -4o available to all freely. Large LLMs trained with trillions of parameters can easily perform masking task effortlessly.)

Name: Muhammad Zohaib Khan Email: zohaibkhanzohaibi2003@gmail.com

Common Errors:

- But smaller LLMs, like gpt2 or EleutherAI/gpt-neo-1.3B, which I used, are inefficient and cannot perform the task of masking correctly. They hallucinate (request access for Llama model was sent, I haven't received access yet)
- Encoder model made few errors in masking uncommon names that didn't belong to hierarchy of names in the dataset. (Ngidi, a name of South African cricketer was not masked correctly).

Potential Improvements:

- Fine-tune models with a more diverse dataset, including varied PII formats (i.e, Address, Contact, Gender etc).
- Use rule-based post-processing to reduce false positives.
- For smaller LLMs, prompt engineering should be performed to get structured outputs and correct masking.

Conclusion:

Encoder models fine tuned on better more variable data and very large LLMs can perform NER tasks accurately.

Meanwhile encoder models trained on linited data perform well only within that data scope and smaller LLMs need efficient prompting get some acceptable results.

Name: Muhammad Zohaib Khan Email: zohaibkhanzohaibi2003@gmail.com