**Evaluation Section of the Report**

In this section, we were trying to find an appropriate evaluation metrics for our project. The primary test data utilized in this analysis includes a Target Email, an Email Generated by GPT 3.5, a Target Email being translated twice, Email Title, and a character "X". Our evaluation metrics consist of BLEU, BERT, and ROUGE scores.

**Test Data Overview:**

The test data comprises:

1. **Target Email:** The original email used as the standard for comparison.
2. **Email Generated by GPT 3.5:** An email created by the GPT 3.5 model.
3. **Target Email Translated by DEEPL:** The target email being translated to Chinese and back by the DEEPL service.
4. **Email Title:** The subject line of the target email.
5. **X:** A control variable representing non-relevant text.

**Evaluation Metrics:**

The following metrics were used:

* **BLEU (Bilingual Evaluation Understudy):** Measures the correspondence of phrases between the generated and the target texts.
* **BERT (Bidirectional Encoder Representations from Transformers):** Evaluates the semantic similarity of text representations.
* **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Focuses on the overlap of n-grams between the compared texts.

**Evaluation Results:**

| **Data Type** | **BLEU** | **BERT** | **ROUGE** |
| --- | --- | --- | --- |
| Target | 1.0 | 1.0 | 1.0 |
| GPT3.5 | 0.052389 | 0.843668 | 0.316623 |
| Translated Twice | 0.564568 | 0.96876 | 0.819048 |
| Title | 0.000041 | 0.954038 | 0.140351 |
| X | 0.0 | 0.774057 | 0.0 |

The evaluation results are summarized in the table below:

**Discussion:**

* **BLEU Score Analysis:** BLEU scores followed our expectations in terms of ordering the data types by relevance. However, the numerical differences between scores are minimal, lacking a clear linear pattern, which led us to dismiss BLEU as a primary evaluation metric.
* **BERT Score Analysis:** Despite BERT scores indicating a high semantic similarity across most text types, including the non-relevant "X" text (score of 0.774057), this metric did not effectively differentiate between the nuances of the text types involved. Consequently, BERT was ruled out as a useful metric for this evaluation.
* **ROUGE Score Analysis:** ROUGE scores proved the most effective in capturing the qualitative differences among the texts. The scores aligned well with our expectations both in terms of order and numerical differentiation, making ROUGE the chosen metric for evaluating our project.

Based on the comparative analysis of the metrics, the **ROUGE score** was selected as the most suitable evaluation metric for our project. It consistently reflected the expected trends and displayed significant numerical distinctions appropriate for evaluating the performance of the generated and translated emails.

**Findings Section of the Report**

This section delves into the results derived from various testing phases of our final project, focusing on the training period, manual testing, and the overall conclusions drawn from the evaluation data.

**1. Training Data Analysis**

Throughout the training period, several metrics were monitored to evaluate the performance improvements of our model across multiple epochs. The primary focus was on the evaluation loss and ROUGE scores as indicators of the model's ability to generate text closely matching the target email.

The final perplexity is 14.17, and there is a significant reduction in evaluation loss from 4.3 to 2.65, suggesting the model was learning effectively. However, the ROUGE scores remained low, around 0.04, highlighting a weak correlation with the target content. The lack of improvement in ROUGE scores points to potential issues either with the model's suitability for this specific text generation task or with the dataset's alignment with our objectives.

**2. Manual Testing Results**

To further assess the model's practical performance, we manually tested it by generating 12 emails and rating them on three aspects: relevance, format accuracy, and utility.

**Manual Testing Scores:**

* Relevance to features: 0.25
* Accuracy as an email: 0.42
* Usefulness: 0.25

These scores reveal that the model occasionally produced relevant and potentially useful content but was generally inconsistent and unreliable for practical deployment.

**3. Conclusions from the Findings**

* **Practical Value:** The current model lacks practical value for real-world applications, primarily due to its inability to consistently generate high-quality emails.
* **Decreasing Effectiveness:** Despite a reduction in evaluation loss, the declining ROUGE scores during training suggest diminishing effectiveness, indicating that increasing dataset size alone may not necessarily lead to better performance.
* **Need for Alternatives:** Considering the current model's limitations, there is a strong case for exploring alternative models or making substantial adjustments to the training process or data handling.

Based on these findings, further actions could include reevaluating the dataset, considering a different model architecture, or making critical adjustments to the training regimen to enhance the quality and reliability of the generated text.