Tuning LLM for Spam Aware Email Generation

Presented to:

Dr. Shaikh Anowarul Fattah
Professor
Department of EEE, BUET

Presented by:

Sudipto Pramanik
Student ID: 0424062538
Department of EEE, BUET

Rationale of the Project

Problem Statement:

- Many legitimate emails (scholarships, funding requests, job applications) are wrongly flagged as spam.
- Spam filters rely on patterns, sometimes incorrectly classifying important messages.
- This can lead to **missed opportunities** and **communication failures**.

Limitations of Existing AI-Generations:

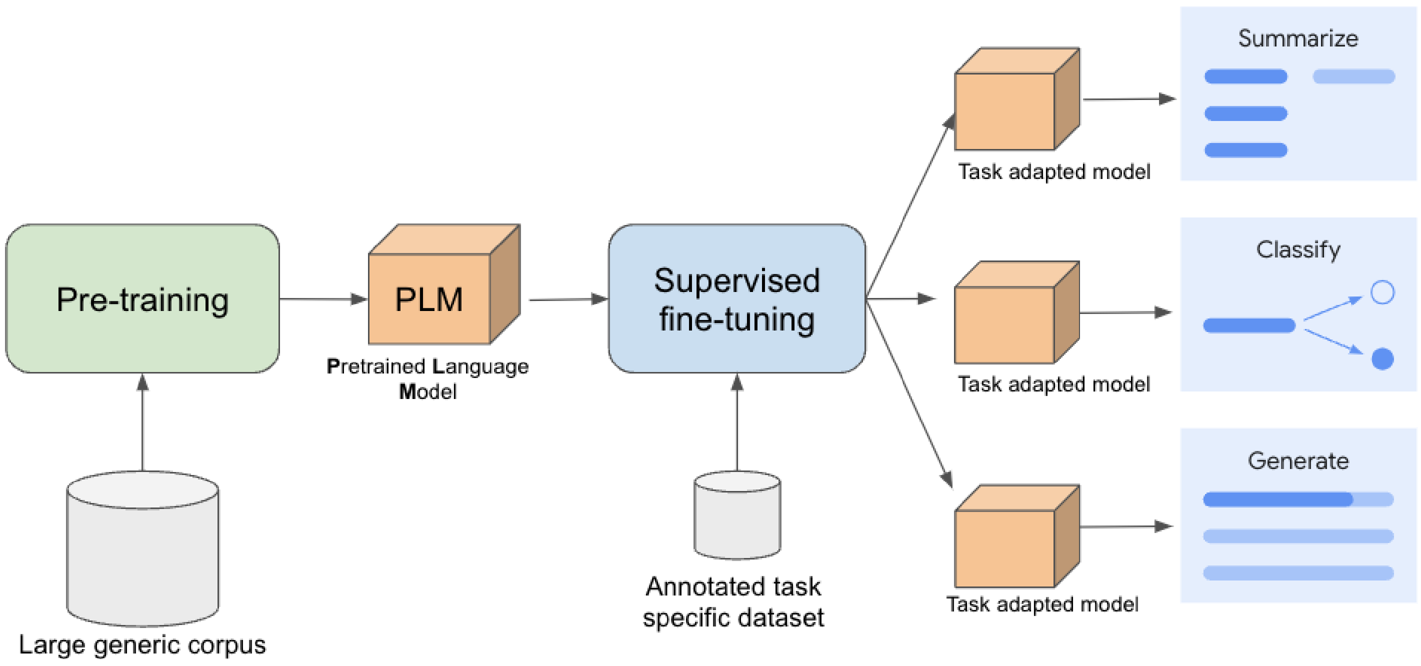
- Conventional AI models generate emails without considering spam filters.
- They may use words/phrases that increase the risk of spam classification.
- No optimization for **deliverability and reliability**.

Project Goal:

- **Develop an AI model that generates professional, non-spam emails.**
- Fine-tune LLMs to avoid spam-triggering patterns.
- Ensure legitimate emails **reach inboxes** without being flagged.

Motivation & Selection of Method Used

• Inspired by *"Controlling Impression: Making ruGPT3 Generate Sentiment-driven Movie Reviews"*, which successfully fine-tuned a model for controlled text generation.
• Supervised Fine-Tuning (SFT) is chosen as it directly optimizes a pre-trained model on labeled datasets.
• The reference paper demonstrates SFT's effectiveness for targeted text generation which has motivated us to use it in our goal of generating non spam email generation.



Dataset

- A Proper Dataset with Mail Statement, Mail Body, Label (Spam/Not Spam) was required for this task.
- A Well-Structured Dataset with these Features was not Available.
- Most of the data, therefore, were generated using GPT Plus after a Well-Defined Prompting Session.
- The Dataset Basically have following Columns: Mail Statement, Mail Content, Label
- This Dataset was Modified and Utilized in Different Tasks as follows:

Mail_Dataset.csv:

- Used for Fine-tuning of LLMs.
- A New Column- “Instruction” was Added. It contains the Necessary Prompt for Fine-tuning.

	A	B	C	D
1	Instruction	Mail Statement	Mail Content	Label

Dataset

TrainTest_Mail.csv:

- To Evaluate the Performance of Fine-tuned Model, a Classifier is trained on this data.
- Two New Columns- “spam” and “not spam” were Added. They are Binary Valued(TRUE/FALSE). For example, “spam” will have the value TRUE if the “Label” is spam.

	A	B	C	D	E	F
1	Instruction	Question	Answer	Label	spam	not spam

Evaluation_Mail.csv:

- Used for Validation of the Classifier.
- Structure: Same as **TrainTest_Mail.csv**.

Unique_Mail_Statements.csv/ Unique_Mail_Statements_Mistral.csv:

- Mail Statements and Generated Responses from Fine-tuned Models and Base Models for Comparison.

	A	B	C
1	Mail Statement	Fine_Tuned_Model_Response	GPT2_Response

Selection of LLMs

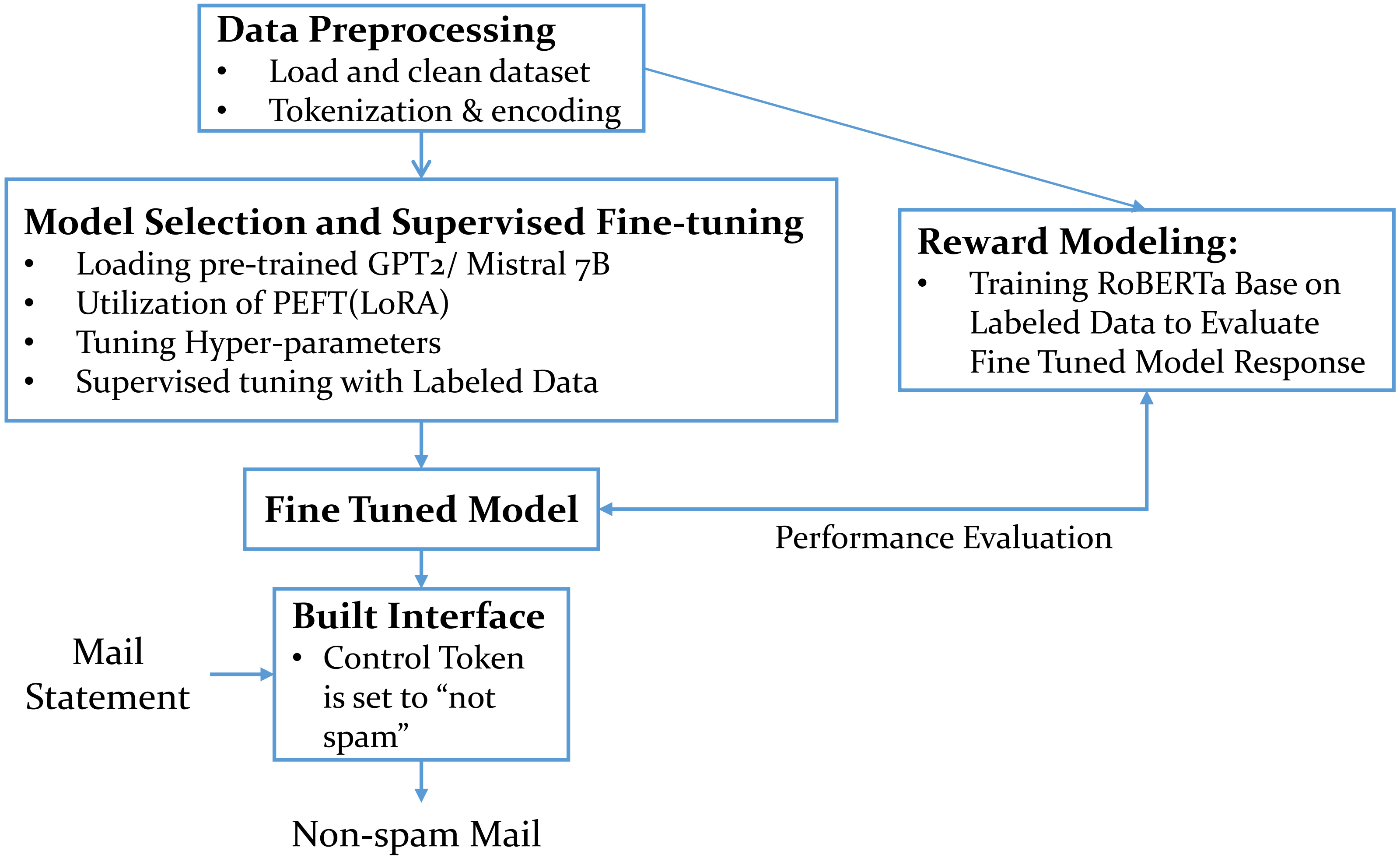
GPT-2

- Developed by OpenAI
- Generate Good-quality Text based on Given Prompt
- Adaptability to Various Types of Questions and Contexts
- Provides the opportunity to test the training method on simpler models.

Mistral-7B

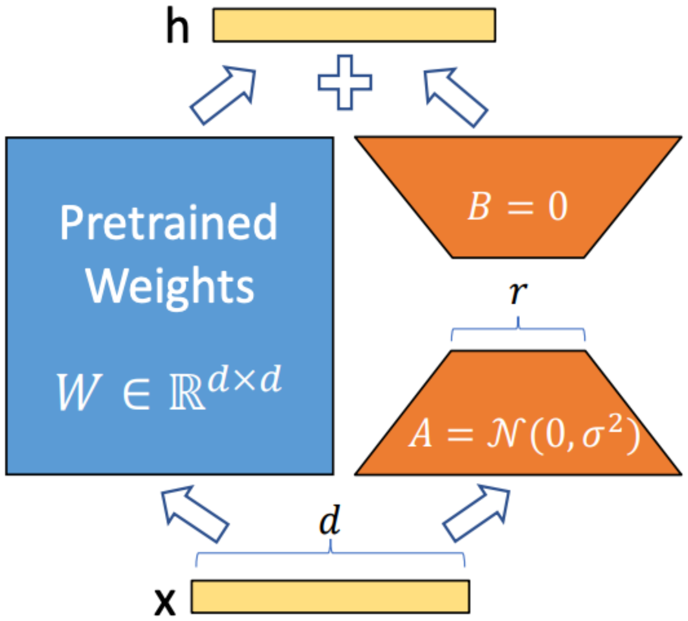
- Developed by Mistral AI
- High Quality and Relevance of Generated Answers
- Can Handle Diverse and Complex Inputs
- Provides the opportunity to test the training method on advanced models.

Methodology- Overall Workflow



Methodology- Setup for Fine-tuning

Parameter Efficient Fine Tuning (PEFT):



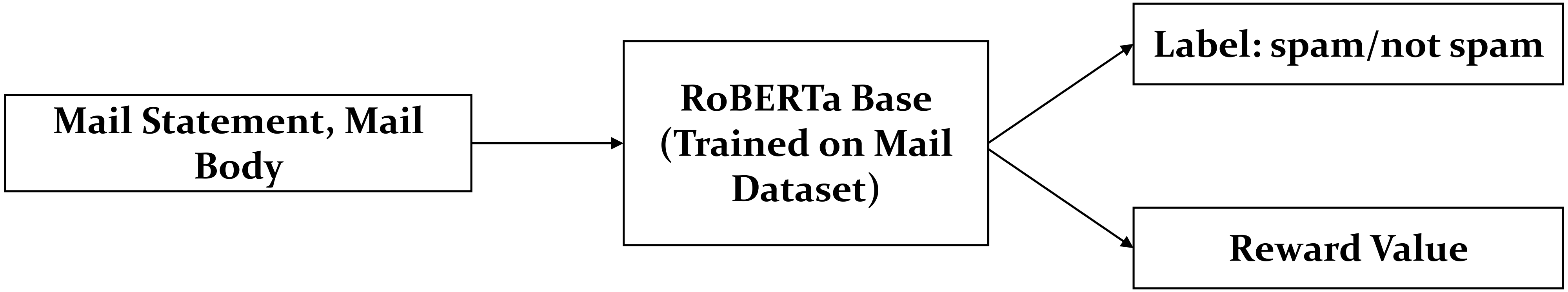
- **LoRA (Low-Rank Adaptation)** optimizes fine-tuning by adding small trainable matrices to transformer attention layers while keeping most model parameters frozen.
- **PEFT (Parameter-Efficient Fine-Tuning)** using LoRA reduces memory usage, speeds up training, and maintains model performance with minimal computational cost.

Methodology- Setup for Fine-tuning

Training Setup & Hyperparameters:

Category	Parameter Name	Value
Model	Base Model	gpt2-large/ mistral 7B
	Fine-Tuning Method	LoRA (Low-Rank Adaptation)
LoRA Parameters	r (LoRA Rank)	16
	lora_alpha (Scaling Factor)	32
	lora_dropout (Dropout Rate)	0.05
Training Parameters	Batch Size	1 (per device)
	Gradient Accumulation Steps	4
	Learning Rate	2e-4
	Optimizer	"paged_adamw_8bit"
	Precision	fp16=True (Half-Precision)
	Max Training Steps	1000

Methodology- Reward Modeling



- Reward Value is taken as the Logit Associated with the Label “not spam”. Thus, a higher reward represents a good response.

Results and Findings

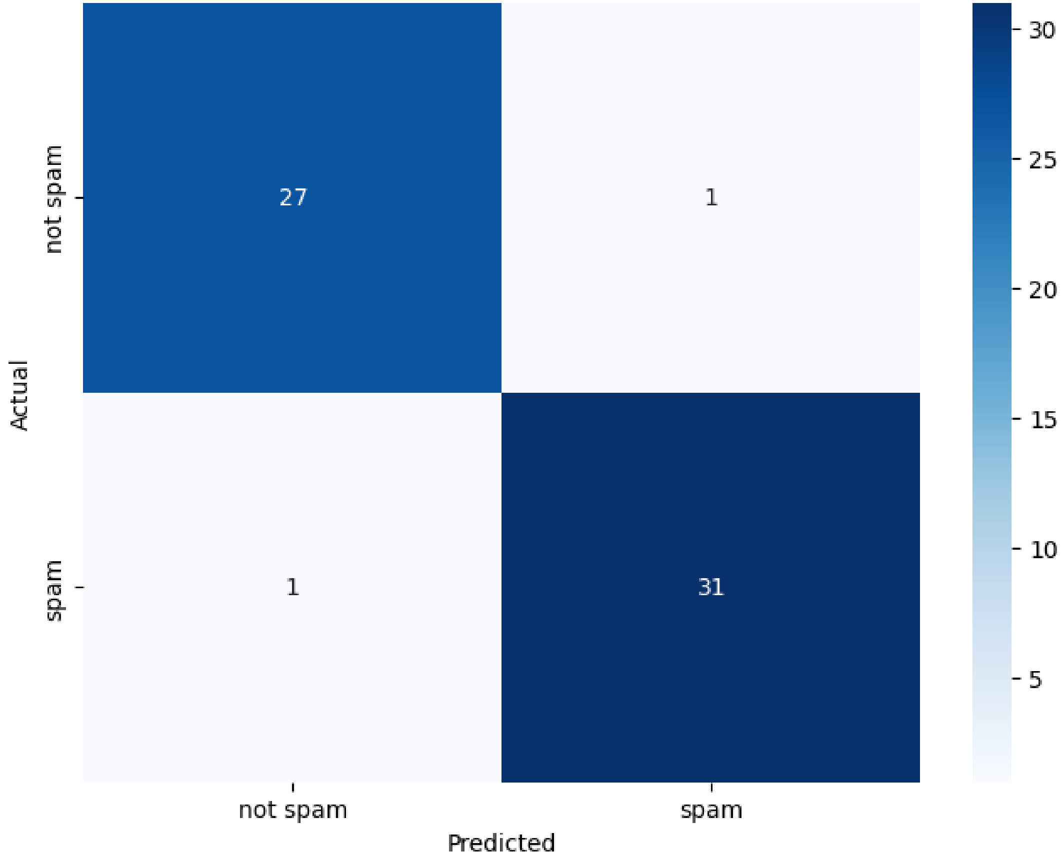
(i) Performance Analysis of Reward Model:

Validation Accuracy: 97%

[6/6 00:02]

```
{'eval_loss': 0.17131340503692627,
 'eval_f1': 0.967032967032967,
 'eval_roc_auc': 0.9670329670329672,
 'eval_accuracy': 0.967032967032967,
 'eval_runtime': 2.6437,
 'eval_samples_per_second': 34.421,
 'eval_steps_per_second': 2.27,
 'epoch': 5.0}
```

Confusion Matrix



Results and Findings

(ii) Fine Tuned Response [Application Interface]:

GPT-2

Mail Prompt

Follow-up on a request for technical support for a mobile app.

Non Spam Mail Generation

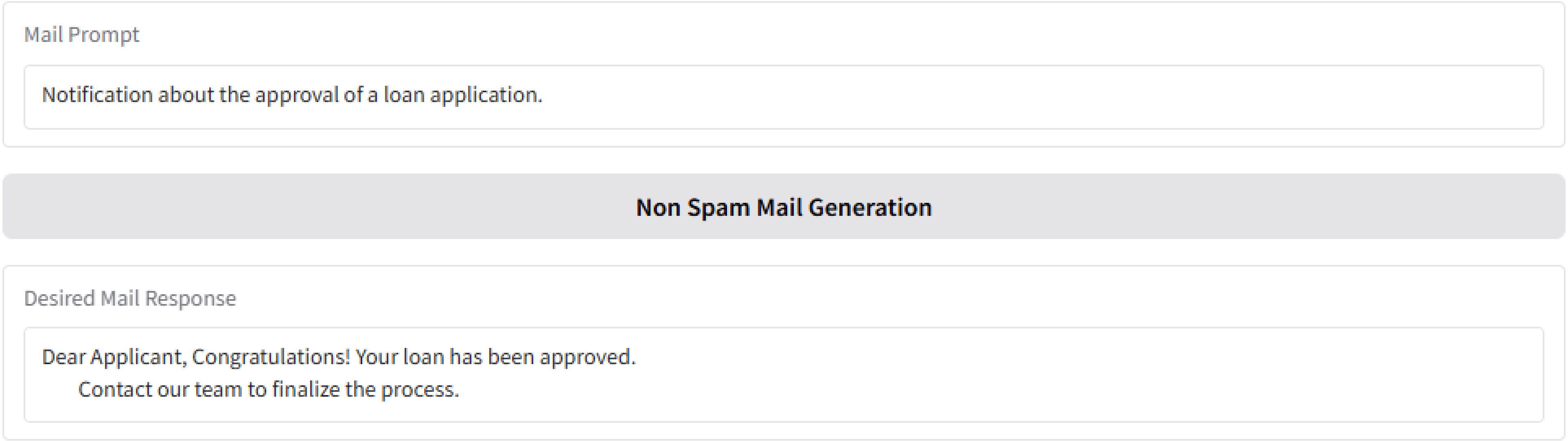
Desired Mail Response

Dear User, Thank you for reaching out to our support team.
We are working on your issue and will update you shortly.

Results and Findings

(ii) Fine Tuned Response [Application Interface]:

<u>Mistral-7B</u>

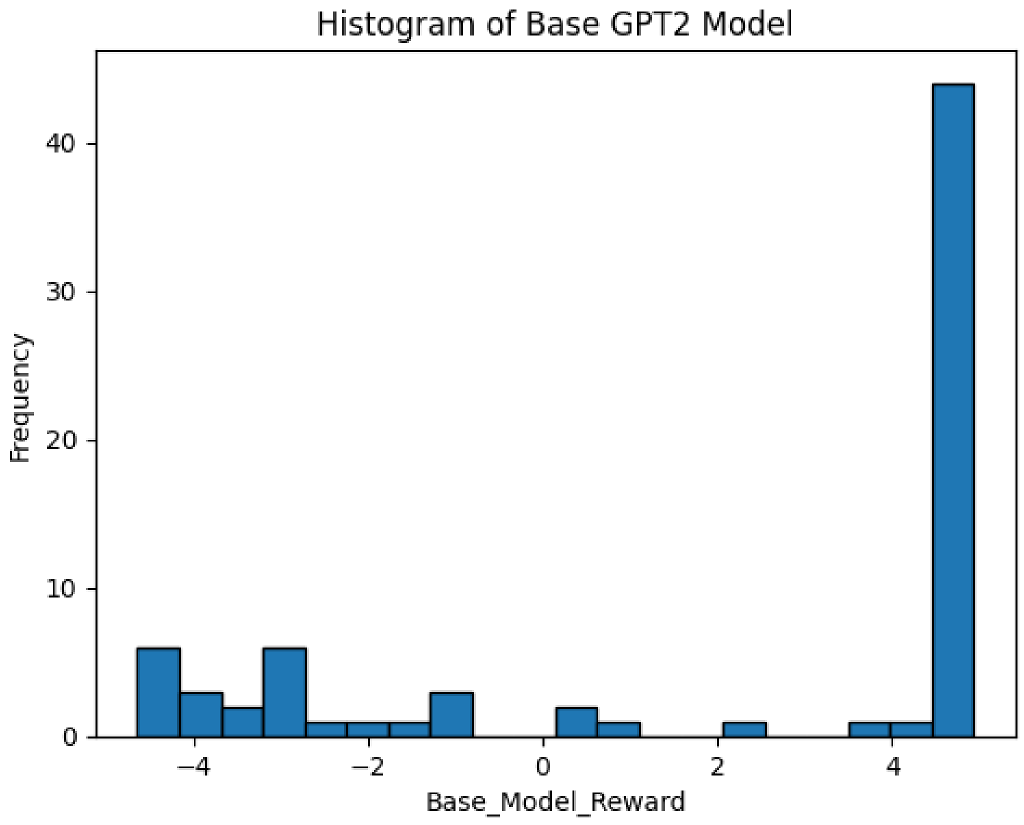


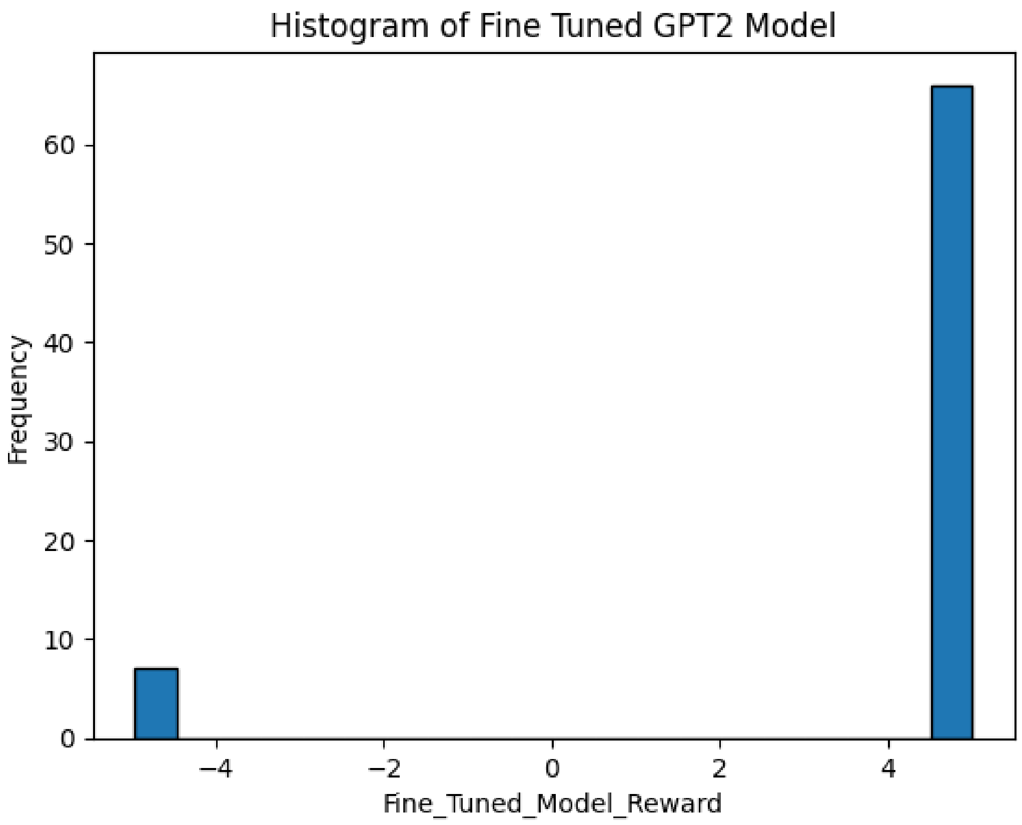
- It is Evident that Built Application is able to Generate Consistent Emails with the Context and Free from any Pattern that will lead it to be flagged as Spam.

Results and Findings

(iii) Comparison between Base Model and Fine-tuned Model:

<u>GPT-2</u>





Results and Findings

(iii) Comparison between Base Model and Fine-tuned Model:

GPT-2

	Base GPT-2	After Fine-tuning
Mean Reward	2.0735743854143847	4.020236119832078
Variance	13.72517049031143	8.646590934661756
% Good Response	68.49%	90.41%

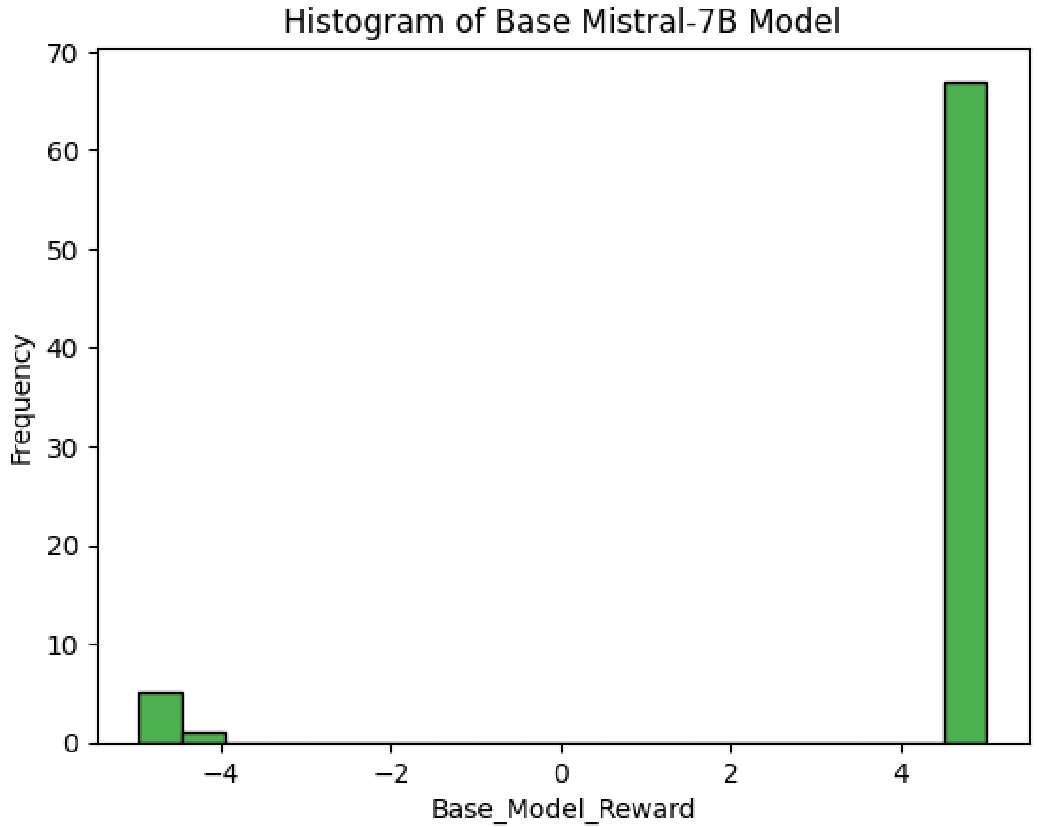
[GPT-2 Reward Value Distribution]

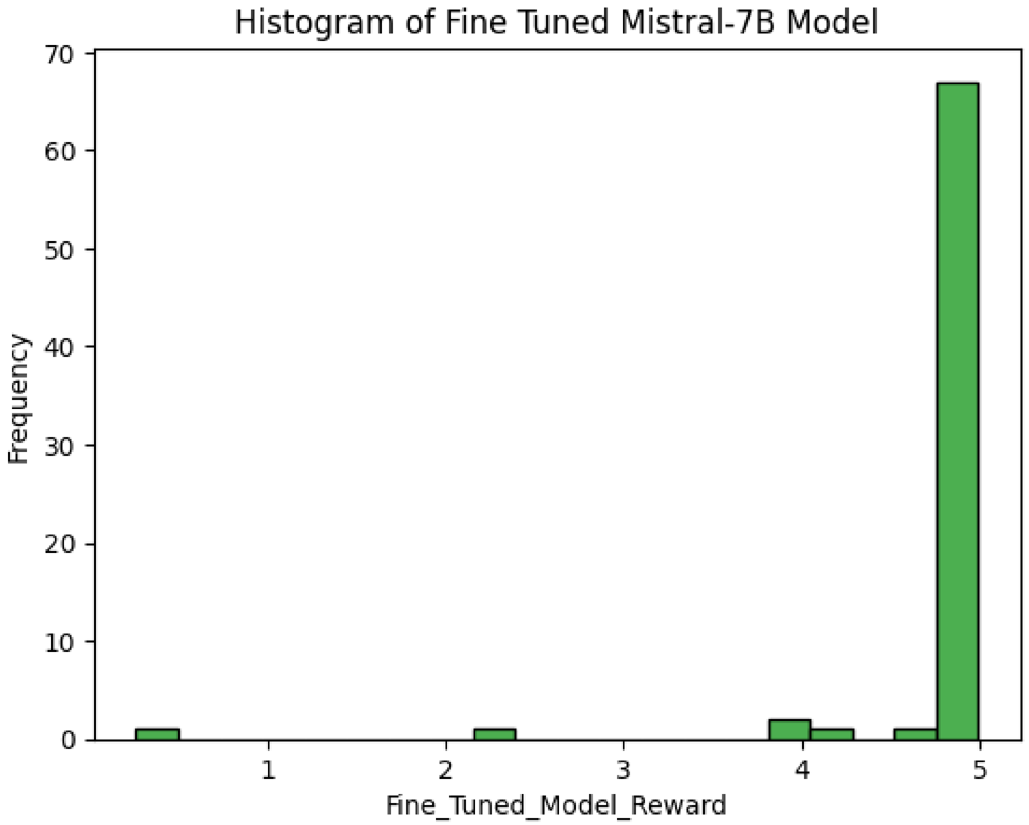
```
Total Rows in Data: 73
Tuning Improves Reward for: 66 Rows
Percentage of Cases Tuning Improves Reward: 90.41095890410958 %
```

Results and Findings

(iii) Comparison between Base Model and Fine-tuned Model:

Mistral-7B





Results and Findings

(iii) Comparison between Base Model and Fine-tuned Model:

Mistral-7B

	Base Mistral 7B	After Fine-tuning
Mean Reward	4.175517545987482	4.804342375226216
Variance	7.367319096872436	0.4224202204245567 4
% Good Response	91.78%	100%

[Mistral 7B Reward Value Distribution]

```
Total Rows in Data: 73
Tuning Improves Reward for: 61 Rows
Percentage of Cases Tuning Improves Reward: 83.56164383561644 %
```

Results and Findings

Key Findings:

- Fine-tuned Models Surpasses Corresponding Base Models.
- Supervised Fine-tuning ensures Consistent Mail Response with the Natural Flow of Writing.
- Our Approach is Applicable both for Simple and Advanced LLMs i.e. GPT-2 and Mistral-7B.
- Same Approach can be used for other Targeted Text Generation Tasks.

Limitations

- This work was conducted using GPT-generated data. In this case, the sample emails were short, which led the fine-tuned models to generate shorter responses. Using a natural dataset could help mitigate this issue.

- The dataset was dominated by data from a few domains. Incorporating more diverse data could enhance the fine-tuned model's performance and generalizability.

References

- Margolina, A. V. (2022). *Controlling impression: Making ruGPT3 generate sentiment-driven movie reviews*. *Journal of Applied Linguistics and Lexicography, 4*(1), 15–25. https://doi.org/10.33910/2687-0215-2022-4-1-15-25
- https://huggingface.co/docs/trl/en/sft_trainer
- https://huggingface.co/openai-community/gpt2-large
- https://huggingface.co/mistralai/Mistral-7B-v0.1

For Reuse and Deployment

- https://huggingface.co/SudiptoPramanik/GPT2_FineTunedModel_for_Non-spam_Mail_Generation
- https://huggingface.co/SudiptoPramanik/Mistral_FineTunedModel_for_Non-spam_Mail_Generation
- https://huggingface.co/SudiptoPramanik/RewardModel_RobertaBase
- https://huggingface.co/spaces/SudiptoPramanik/GPT2_Non_Spam_Email_Generation