An Automatic Spam-Aware Email Generation Scheme Based on Efficient Fine-Tuning of Large Language Models

Sudipto Pramanik[1], Sudha Bhattacharjee[2], Sandipa Chowdhury[1], Shaikh Anowarul Fattah[1], Motasim Billah[1]
[1]Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka, 1000, Bangladesh
[2]Department of Computer Science and Engineering, Daffodil International University, Dhaka, 1216, Bangladesh
Email: pramaniksudipto4@gmail.com, sudhabhattacharjee5@gmail.com, csandipa556@gmail.com, fattah@eee.buet.ac.bd, billahmotasim2605@gmail.com

***Abstract*—With the growing use of AI-driven communication, large language models (LLMs) have become popular tools for automated email generation. However, these models are typically unaware of how spam filters operate, often generating content that inadvertently contains spam-triggering words or patterns. As a result, even legitimate emails, such as job applications or scholarship offers, are frequently misclassified as spam. This misclassification can result in missed opportunities and communication failures. This paper proposes an efficient approach for generating professional, non-spam emails based on fine-tuning an LLM. In order to avoid spam-triggering patterns, supervised fine-tuning (SFT) approach has been incorporated in the proposed method, along with parameter efficient fine-tuning (PEFT) techniques. Here to optimize the pre-trained models, GPT-2 and Mistral-7B models are chosen. From extensive simulations, it is found that the proposed fine-tuned models outperform base models in generating contextually appropriate emails that are less likely to be flagged as spam. The proposed scheme is applicable to both simple and advanced LLMs and can be extended to other targeted text generation tasks.**

***Index Terms*—Spam-aware email generation, large language model (LLM), supervised learning, fine-tuning deep learning models, generative learning models.**

I. Introduction

Email communication is a cornerstone of modern professional and personal interactions, facilitating critical exchanges such as job applications, scholarship notifications, and business correspondence. However, the increasing sophistication of spam filters has introduced a significant challenge: the misclassification of legitimate emails as spam. Recent industry benchmarks report that only 83–85% of legitimate emails reach the inbox, meaning roughly 15–17% are diverted to spam or filtered—effectively functioning as false positives [1]–[3]. This problem is exacerbated by the reliance of traditional spam filters on pattern recognition, which often fails to distinguish between malicious spam and important messages [4]. As a result, there is a pressing need for solutions that ensure the reliable delivery of legitimate emails while maintaining the effectiveness of spam filters.

The development of spam-aware email generation systems is driven by advancements in three key research domains: (1) the evolution of LLMs, (2) fine-tuning techniques for task-specific adaptation, and (3) the design and limitations of spam filters. This section reviews prior works in these areas, emphasizing recent developments and identifying gaps that our research addresses. The LLMs have emerged as powerful tools for natural language generation, enabling applications ranging from conversational agents to content creation. OpenAI's GPT-2 [5] and Mistral AI's Mistral-7B [6] represent significant milestones in this domain. These models have revolutionized the field of natural language processing (NLP), enabling the generation of high-quality text . They leverage transformer architectures to generate coherent and contextually relevant text. Recent work has extended LLMs to specialized domains, such as healthcare [8] and legal document drafting [9], demonstrating their versatility. However, these models are not inherently designed to account for the nuances of spam classification. Emails generated by conventional LLMs often contain phrases or structures that trigger spam filters, reducing their deliverability [16]. This limitation underscores the need for models that are optimized for both content quality and inbox placement.

Traditional spam filters rely on pattern recognition and machine learning algorithms to classify emails as spam or non-spam [4].Recent research has focused on improving the accuracy of spam filters using advanced techniques like deep learning [14] and reinforcement learning [15]. However, these approaches primarily focus on improving the classification accuracy of spam filters rather than optimizing the content generation process to avoid spam triggers.

Fine-tuning a pre-trained model for specific tasks has become a cornerstone of modern NLP research. Different fine-tuning techniques are getting popularity to adapt LLMs for tasks, such as sentiment analysis, text classification, and controlled text generation [10]. Among these methods, low-rank adaptation (LoRA) [11] and adapter modules [12] have further enhanced the efficiency of fine-tuning by reducing memory usage and computational costs. These techniques enable the adaptation of large models like GPT-2 and Mistral-7B for specific tasks without requiring extensive retraining

[13]. Despite these advancements, the application of fine-tuning techniques to spam-aware email generation remains underexplored.

While significant progress has been made in the development of LLMs, fine-tuning techniques, and spam filters, there is a notable gap in research that bridges these areas. Existing work has primarily focused on improving the quality of text generation or the accuracy of spam classification, but few studies have addressed the challenge of generating emails that are both contextually appropriate and less likely to be flagged as spam [16]. Recent studies have explored the use of LLMs for email generation [7], [17], but these efforts have not explicitly addressed the issue of spam classification.

The major challenge here is to bridge the gap between AI-generated content and real-world usability. By incorporating scoring model and efficient fine-tuning techniques, we aim to create a system that generates professional, non-spam emails while maintaining computational efficiency. The fine-tuning scheme is designed by combining a supervised fine-tuning (SFT) and a parameter-efficient fine-tuning (PEFT) techniques that optimizes LLMs for spam-aware email generation. By fine-tuning pre-trained models like GPT-2 and Mistral-7B on a carefully curated dataset, we aim to generate emails that are not only contextually appropriate but also less likely to be flagged as spam.

The contributions of this paper are threefold:

- A spam-aware email generation framework is proposed by fine-tuning pre-trained models like GPT-2 and Mistral-7B on a carefully curated dataset.
- The fine-tuning scheme is designed by combining SFT and PEFT techniques that optimizes LLMs for spam-aware email generation.
- We demonstrate the effectiveness of our approach through extensive experiments, showing that fine-tuned models outperform base models in generating contextually appropriate spam aware emails.

II. Methodology

This work seeks to improve language models for spam-aware email generation. The systematic process to achieve this target is outlined in Figure 1, which illustrates the key steps, including dataset preparation, data processing, supervised fine-tuning of language model, reward modeling, and evaluation of performance.

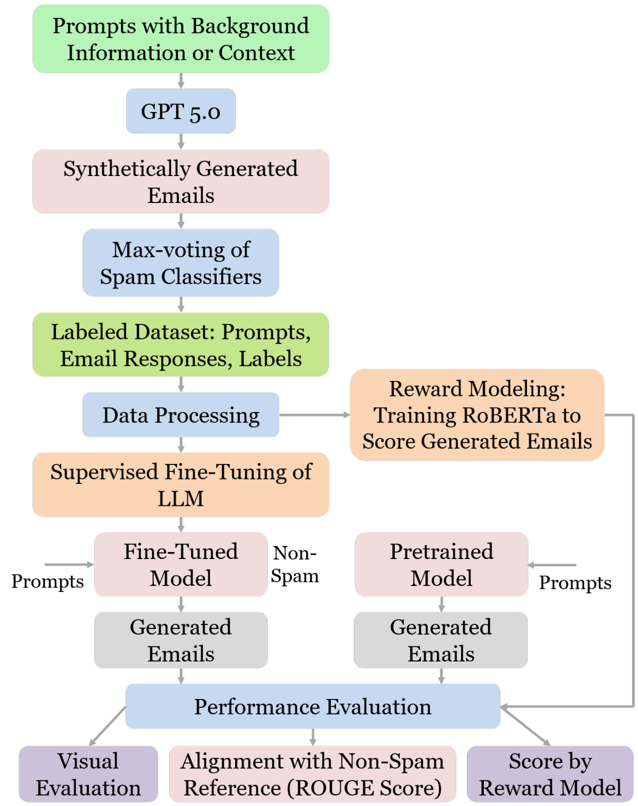


Fig. 1: Proposed Framework for Spam-Aware Email Generation

TABLE I: Samples from the Custom Dataset

Prompt	Email Response	Label
Confirm a client meeting with Mr. Karim for ACME project on Monday 10:00…	Dear Mr. Karim, I am writing to confirm our project update meeting on next…	Non Spam
Write an email to Mr. Abid notifying him about the application fee waiver granted for him; include deadline and contact details…	Congratulations! You have been selected for an exclusive fee waiver at our prestigious university. Click here to confirm immediately…	Spam
Write a follow up on scholarship application submitted four weeks ago to Florida…	Dear Admission Committee, Greetings! I applied for a scholarship on last month…	Non Spam

A. Dataset Preparation

In order to tune the LLMs for generating professional emails that are less likely to be flagged as spam, a labeled dataset containing email prompts, generated mail responses, and corresponding labels was required. However, existing corpora such as the Enron Email Dataset and the SpamAssassin corpus do not provide data in this structured form. Also, because the goal is to address the inherent spam patterns present in emails generated by language models, it was expected to build the dataset using synthetically generated emails. Accordingly, GPT-5.0 was employed to generate 500 prompt–email pairs through a carefully designed prompting session that captured both legitimate and spam-triggering attributes. These synthetic emails were then passed through a max-voting ensemble of three high-performing spam classifiers from Hugging Face—*dima806/email-spam-detection-roberta* (98.7% accuracy), *mariagrandury/roberta-base-finetuned-sms-spam-detection* (99.8% accuracy), and *dima806/email-spam-detection-distilbert* (99.2% accuracy)—to assign spam/not-spam labels. The resulting labeled dataset of prompts, responses, and labels was subsequently processed and used in the next stages of our framework. A few samples from the dataset are shown in Table I.

B. Data Processing

After constructing the custom email dataset, we performed a careful processing stage to ensure consistency and readiness for training. Blank or incomplete entries, duplicate prompts, and minor formatting errors were removed to prevent errors during model training. Following this cleaning step, punctuation and casing were standardized so that all prompts, responses, and labels followed a uniform structure. These steps produced a clean, well-organized dataset suitable for fine-tuning the LLMs and for generating reliable spam-aware email outputs.

C. Supervised Fine-Tuning (SFT)

The pretrained LLM (GPT-2 or Mistral-7B) was fine-tuned on our curated prompt–response pairs annotated with spam/non-spam labels. Although the base models exhibit strong general language understanding, they are not inherently optimized to avoid spam-triggering patterns in generated emails. In our setup, supervised fine-tuning conditions the model on the label associated with each example so that it learns to generate outputs consistent with the desired class. During inference, the label can be fixed to "non-spam," enabling the model to reliably produce professional, spam-aware emails that maintain proper structure, tone, and contextual relevance.

To make this adaptation efficient, we used parameter-efficient fine-tuning techniques that update only a small subset of parameters instead of the entire model. Specifically, LoRA was used to apply low-rank weight updates, and QLoRA incorporated quantization to further cut memory requirements.

The fine-tuning was driven by a cross-entropy loss function that reduced the difference between the generated and target responses, thereby improving the quality of spam-aware email generation.

D. Reward Modeling

To quantitatively evaluate the effectiveness of the fine-tuned LLMs, we employed a RoBERTa-base classifier as a reward model. This model was trained on our labeled dataset to distinguish between spam and non-spam emails and is used solely for evaluation purpose.

For each prompt p_i, both the base model and the fine-tuned model generate a response $y_i^{(m)}$, where $m \in \{\text{base}, \text{fine}\}$. The reward model outputs logits $\mathbf{z}_i^{(m)} = (z_{i,\text{spam}}^{(m)}, z_{i,\text{non-spam}}^{(m)})$ for the two classes. The **reward value** for a generated email is defined as the logit corresponding to the non-spam class:

$$R^{(m)}(i) = z_{i,\text{non-spam}}^{(m)}.$$

This reward value and its distribution across all prompts is used to compare the generations of base and fine-tuned models. For N prompts, the mean reward score is

$$\bar{R}^{(m)} = \frac{1}{N}\sum_{i=1}^{N} R^{(m)}(i), \quad \Delta\bar{R} = \bar{R}^{(\text{fine})} - \bar{R}^{(\text{base})},$$

which directly measures improvement in spam-awareness after fine-tuning.

For user-facing visualization, the same logits are converted into non-spam probabilities via a softmax (or sigmoid if only one logit is output):

$$P_{\text{non-spam}}^{(m)}(i) = \frac{\exp\big(z_{i,\text{non-spam}}^{(m)}\big)}{\exp\big(z_{i,\text{spam}}^{(m)}\big) + \exp\big(z_{i,\text{non-spam}}^{(m)}\big)} \in (0,1).$$

These probabilities are displayed alongside each generated email in the application interface so that end users can see, in real time, the estimated likelihood of their message reaching the recipient's inbox without being flagged as spam.

E. Performance Evaluation

The effectiveness of the proposed framework is assessed using three complementary metrics:

- **Visual Evaluation:** Representative outputs from the base and fine-tuned models are examined side by side to qualitatively assess improvements in coherence, contextual relevance, and avoidance of spam-triggering patterns.
- **Score by Reward Model:** The RoBERTa-based reward model assigns a reward value to each generated email. The distribution of these values across all prompts is compared between the base and fine-tuned models to provide a quantitative measure of spam-awareness improvement.
- **ROUGE Score:** Each generated email corresponding to a test prompt is compared with corresponding non-spam reference email using ROUGE metrics to quantify how closely the output aligns with professional, non-spam wording. Responses from both the fine-tuned and pretrained models are evaluated against the reference responses, and the distributions of ROUGE-1, ROUGE-2, and ROUGE-L scores are reported to capture the improvements.

III. Experimental Setup

This section describes the experimental framework used to implement and evaluate the proposed spam-aware email generation system.

A. Dataset and Model Setup

The curated spam-aware email dataset was divided into an 80% training set and a 20% test set. This split was applied both for fine-tuning the LLMs and for training/validating the reward model to ensure fair comparison and evaluation.

Two pre-trained LLMs — **GPT-2 Large** and **Mistral-7B** — were used as base models. Both were fine-tuned using **LoRA (Low-Rank Adaptation)**, which efficiently adapts only a small subset of parameters, preserving model quality while minimizing computational overhead.

B. Fine-Tuning Configuration

The LoRA-based fine-tuning of the LLMs uses the following hyperparameters:

- **LoRA Rank (r):** 16
- **Scaling Factor (lora_alpha):** 32
- **Dropout Rate (lora_dropout):** 0.05

- **Batch Size:** 1 per device
- **Gradient Accumulation Steps:** 4
- **Learning Rate:** 1e-4
- **Optimizer:** `"paged_adamw_8bit"`
- **Precision:** fp16 (Half-Precision)
- **Max Training Steps:** 1000

C. Reward Model Training Configuration

The **RoBERTa-base** classifier is trained separately on the same 80/20 split to serve as the reward model for evaluation. Its training configuration is:

- **Batch Size:** 8 per device
- **Learning Rate:** 2e-5
- **Max Epochs:** 10
- **Optimizer:** AdamW with weight decay 0.01
- **Scheduler:** Linear warmup for first 10% of steps
- **Loss Function:** Cross-entropy

IV. Results and Analysis

This section presents the evaluation of the proposed spam-aware email generation system through qualitative and quantitative analyses of model outputs.

A. Visual Evaluation of Generations

To qualitatively assess the effectiveness of the proposed framework, representative outputs from the pretrained and fine-tuned models were examined side by side. This visual evaluation highlights how the fine-tuned LLMs generate more professional and spam-aware emails compared to their pretrained counterparts.

Figure 2 presents examples of email responses produced by the pretrained GPT-2 model versus the fine-tuned GPT-2 model. While the pretrained GPT-2 occasionally includes overly promotional or awkwardly structured phrases that may trigger spam filters, the fine-tuned GPT-2 consistently produces coherent, contextually appropriate responses with a professional tone and minimal spam-triggering patterns.

Similarly, Figure 3 illustrates the comparison between the pretrained and fine-tuned Mistral-7B models. Although the pretrained Mistral-7B is stronger than GPT-2 in terms of fluency, its responses sometimes contain exaggerated claims or embedded links resembling marketing content. After fine-tuning with the curated dataset, the model's outputs become more concise and aligned with the intended prompt while avoiding such spam-like structures.

B. Reward-Based Evaluation

The RoBERTa-base reward model achieved a validation accuracy of **96.67%**, demonstrating its effectiveness in distinguishing between spam and non-spam emails. The classification performance is illustrated in Table II, while the confusion matrix in Figure 4 further confirms the model's reliability.

Using this reward model, we evaluated the outputs of both the pretrained and fine-tuned LLMs. For each generated email, the reward value was defined as the logit corresponding to the non-spam class. Figures 5a and 5b show the distributions of reward scores for GPT-2 generations before and after fine-tuning, respectively. A clear rightward shift in the distribution indicates that the fine-tuned model produces outputs with higher non-spam tendency. In fact, reward scores improved in approximately 90% of the test prompts for GPT-2.

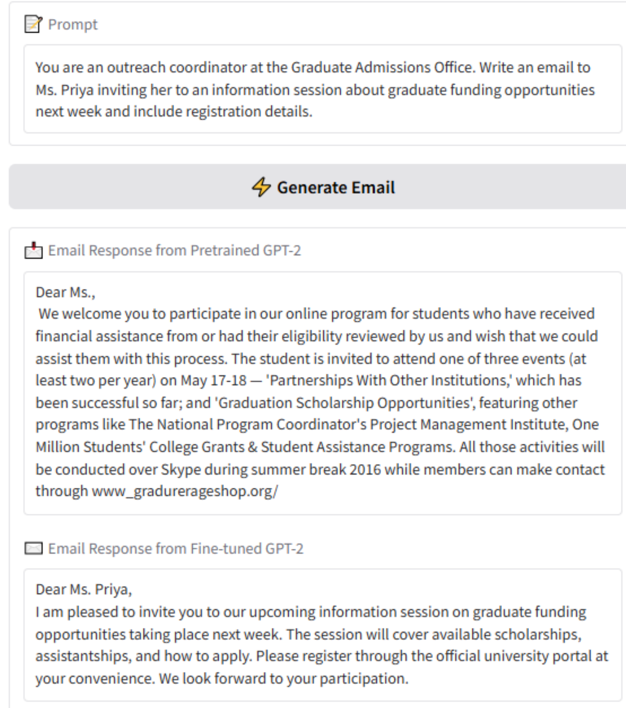


Fig. 2: Visual Comparison between Pretrained and Fine-tuned GPT-2 Generations

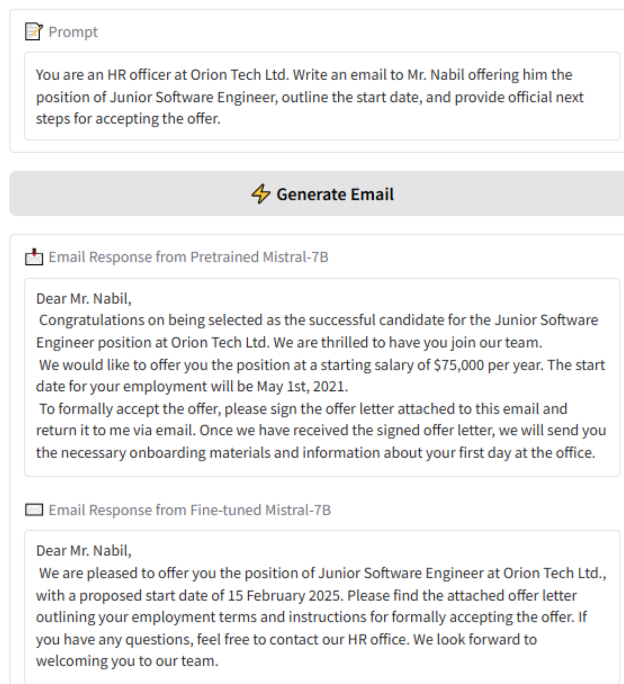


Fig. 3: Visual Comparison between Pretrained and Fine-tuned Mistral-7B Generations

TABLE II: Reward Model Performance on Validation Set

Metric	Value
Accuracy	96.67%
Precision	96.88%
Recall (Sensitivity)	96.88%
Specificity	96.43%
F1 Score	96.88%

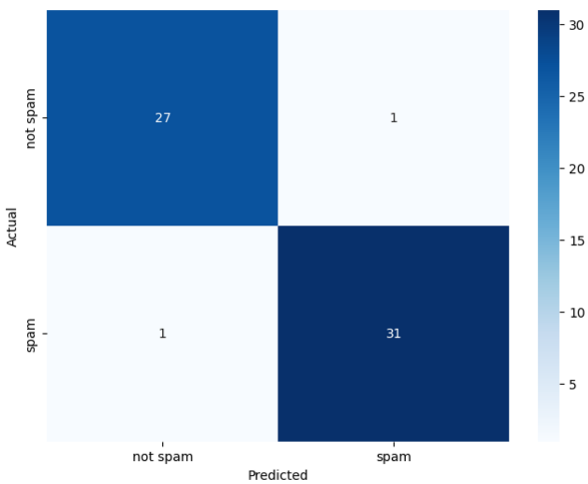


Fig. 4: Confusion Matrix for the Reward Model

A similar trend is observed for Mistral-7B (Figures 6a and 6b), where reward scores increased for about 84% of the test prompts after fine-tuning. Table III summarizes these results, showing higher mean reward scores, lower variance, and a larger proportion of good (non-spam) responses for the fine-tuned models compared to their pretrained counterparts.

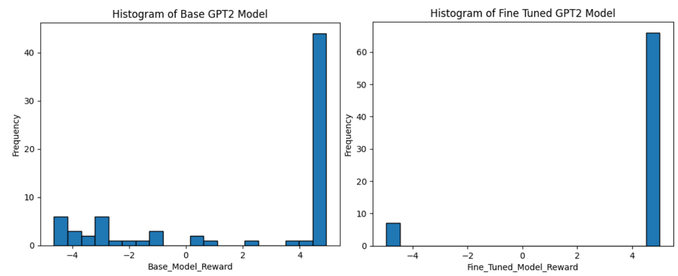
TABLE III: Reward Based Evaluation Summary

Metric	GPT-2		Mistral 7B	
	Pretrained	Finetuned	Pretrained	Finetuned
Mean Reward	2.0736	4.0202	4.1755	4.8043
Variance	13.7252	8.6466	7.3673	0.4224
% Good Response	68.49%	90.41%	91.78%	100%

C. ROUGE Similarity to Non-Spam References

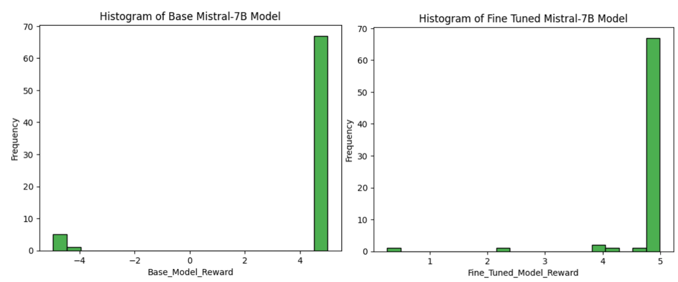
To further assess how closely the generated emails align with professional, non-spam wording, we computed ROUGE scores for each response against its corresponding non-spam reference email in the test set. This metric complements the reward-based evaluation by quantifying textual similarity to high-quality reference emails rather than relying solely on the reward model.

Specifically, ROUGE-1, ROUGE-2, and ROUGE-L scores were computed for the outputs generated by both the pre-trained and fine-tuned models. Figures 7 and 8 present the distributions and spreads of these scores for GPT-2 and Mistral-7B, respectively. The clear rightward shift and higher mean values observed in the fine-tuned models demonstrate that their generated emails more effectively avoid spam-triggering patterns and exhibit a professional tone and structure closely matching the reference emails.



(a) Pretrained GPT-2 (b) Fine-tuned GPT-2

Fig. 5: Reward Score Distributions for GPT-2



(a) Pretrained Mistral-7B (b) Fine-tuned Mistral-7B

Fig. 6: Reward Score Distributions for Mistral-7B

D. Application Interface

An interactive interface was built using `Gradio` to demonstrate the proposed system (Figure 9). This allows users to quickly generate professional, spam-aware emails and view the estimated inbox placement likelihood.

V. Conclusion and Future Works

This paper proposed an efficient fine-tuning scheme for large language models to generate professional, spam-aware emails. By combining supervised fine-tuning with parameter-efficient techniques such as LoRA and QLoRA, GPT-2 and Mistral-7B were adapted to consistently produce contextually appropriate emails while avoiding spam-triggering patterns. Extensive experiments using reward scores, ROUGE similarity and visual comparisons showed clear improvements over the pretrained models in tone, coherence and deliverability. The approach proved effective across both smaller and larger models, demonstrating its versatility. In future work, we aim to expand the dataset with more diverse emails, explore richer feedback signals, and extend the framework to other targeted text generation tasks, such as customer support or automated business correspondence, where maintaining context and avoiding spam classification is crucial.

References

[1] Validity, "2020 Email Deliverability Benchmark Report," *Validity Inc., Industry Benchmark Report*, Boston, MA, USA, 2020. [Online]. Available: https://www.validity.com/wp-content/uploads/2020/03/2020-Email-Delivery-Benchmark.pdf

[2] Validity, "2023 Email Deliverability Benchmark Report," *Validity Inc., Industry Benchmark Report*, Boston, MA, USA, 2023. [Online]. Available: https://www.validity.com/wp-content/uploads/2023/03/2023-Email-Deliverability-Benchmark.pdf

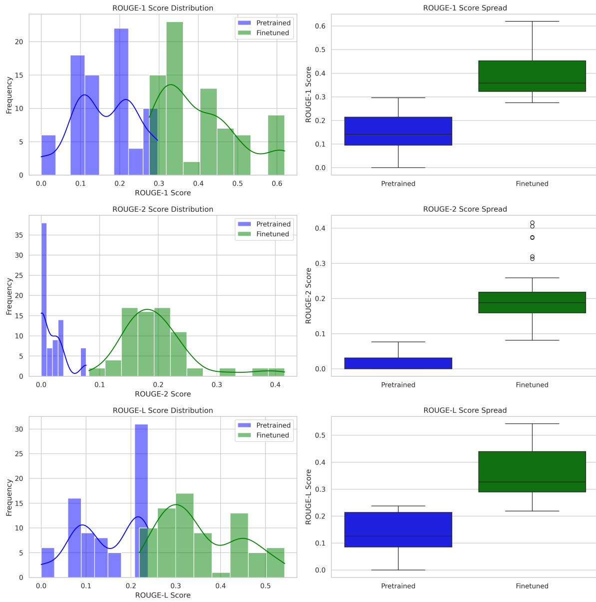


Fig. 7: ROUGE Score Distributions (left) and Spreads (right) for Pretrained and Fine-tuned GPT-2

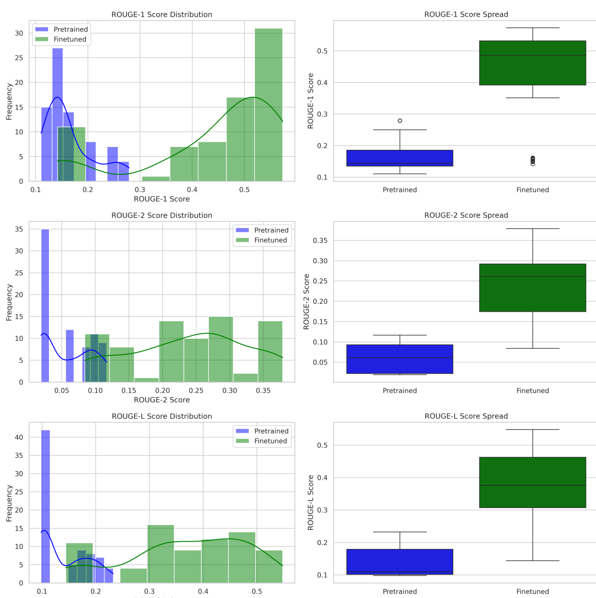


Fig. 8: ROUGE Score Distributions (left) and Spreads (right) for Pretrained and Fine-tuned Mistral-7B

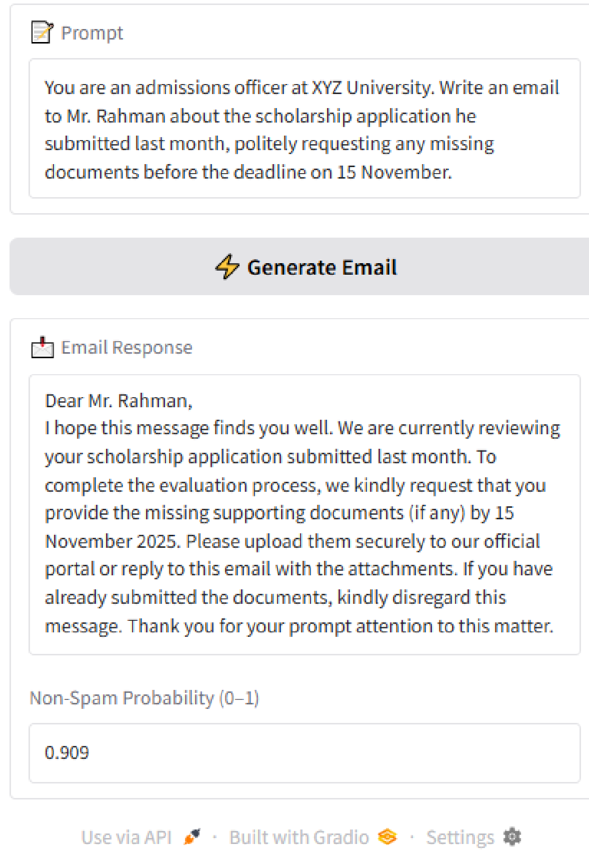


Fig. 9: Built Application Interface

[3] EmailToolTester, “Email Deliverability Statistics,” *EmailToolTester, Industry Benchmark Report*, 2024. [Online]. Available: https://www.emailtooltester.com/en/blog/email-deliverability-statistics/

[4] G. V. Cormack, “Email Spam Filtering: A Systematic Review,” *Foundations and Trends in Information Retrieval*, vol. 1, no. 4, pp. 335–455, 2008. [Online]. Available: https://doi.org/10.1561/1500000006

[5] OpenAI, “GPT-2: Language Models are Unsupervised Multitask Learners,” 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language$_{m}odels_{a}re_{u}nsupervised_{m}ultitask_{l}earners.pdf$.

[6] Mistral AI, “Mistral-7B: High-Quality Text Generation,” 2023. [Online]. Available: https://mistral.ai/technical-paper.

[7] J. Smith, A. Kumar, and F. Li, “Prompting for Email Drafting: Can LLMs Aid in Professional Email Composition?” in *Proc. of NAACL/EMNLP*, 2024.

[8] K. Singhal *et al.*, “Large language models in medicine,” *Nature Medicine*, vol. 29, no. 8, pp. 1930–1940, 2023. [Online]. Available: https://doi.org/10.1038/s41591-023-02448-8

[9] N. Guha *et al.*, “LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models,” in *Proc. NeurIPS Datasets and Benchmarks Track*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2308.11462

[10] J. Howard and S. Ruder, “Universal Language Model Fine-tuning for Text Classification,” *arXiv preprint arXiv:1801.06146*, 2018.

[11] E. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” *arXiv preprint arXiv:2106.09685*, 2021.

[12] N. Houlsby et al., “Parameter-Efficient Transfer Learning for NLP,” in *Proceedings of the International Conference on Machine Learning*, pp. 2790–2799, 2019.

[13] Y. Liu et al., “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,” *arXiv preprint arXiv:2107.13586*, 2021.

[14] G. Nasreen *et al.*, “Email spam detection by deep learning models using novel feature selection,” *Information Sciences*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110866524000367

[15] X. Wang *et al.*, “Spam Filtering in the Modern Era: A Review of Machine & Deep Learning Models,” in *Proc. of the 2nd International Conference on Data Analysis and Machine Learning (DAML)*, 2024. 2025. [Online]. Available: https://www.scitepress.org/Papers/2024/135260/135260.pdf

[16] N. Bouchareb and I. Morad, “Analyzing the Impact of AI-Generated Email Marketing Content on Email Deliverability in Spam Folder Placement,” *Holistica Journal of Business and Public Administration*, vol. 15, no. 1, pp. 96–106, 2024. [Online]. Available: https://doi.org/10.2478/hjbpa-2024-0006

[17] D. Grewal, A. Roggeveen, J. Nordfält, and K. Lemon, “How Generative AI Is Shaping the Future of Marketing,” *Journal of the Academy of Marketing Science*, vol. 52, pp. 1097–1116, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s11747-024-01064-3