Published by The College of Science & Information Technology (COSIT), TASFUED, Vol. 20, No. 1, pp. 1-14..

COSIT, TASFUED
Journal of Science and Information Technology (JOSIT)

A Fine-Tuned GPT-2 For Intelligent Auto-Response System Using Transfer Learning with Contrived Email Data

[1]Osofuye, O.; [1]Sogunle, O.; [2]Osofuye A.; [1]Ogunyale O.; [1]Jegede T.; [1]Abiodun, T.;

[1]Oni. A.

[1]Department of Computer and Information Sciences, Covenant University, Ota, Nigeria.
[2]Business Risk and Internal Control, CapitalSage Holdings, Nigeria.

Corresponding Author: odunayo.osofuye@covenantuniversity.edu.ng
Other Authors: olubusolami.sogunle@stu.cu.edu.ng; a.osofuye@capitalsage.ng; ogunyale.oluwaseun@covenantuniversity.edu.ng; tope.jeged@covenantuniversity.edu.ng; theresa.abiodun@covenantuniversity.edu.ng; aderonke.oni@covenantuniversity.edu.ng

Abstract

This study involves improving intelligent auto-response systems by fine-tuning GPT-2 using a contrived email dataset through transfer learning. Existing natural language models lacks the contextual specificity and understanding necessary for generating coherent and relevant email responses. Initial experiments using the standard Enron corpus revealed that general-purpose datasets often lack the contextual specificity required for high-quality email generation. To tackle this gap, GPT-2 was fine-tuned on a specialised dataset developed to mimic diverse email interactions to adapt its generative capabilities for various email contexts, such as professional communication, non-academic questions, and academic inquiries. The results of experiments show significant improvements with an increase in BLEU score from 0.53 to 0.66 compared to the baseline model, in response quality. Also, the contrived fine-tuned model achieved substantial gains in ROUGE scores across all variants indicating better contextual understanding and relevance (e.g., ROUGE-1 F1: 0.580) over the baseline. The human evaluations and qualitative analysis conducted support these findings, confirming the ability of the model to generate relevant replies that is largely close to expected human behaviour. The potential of transfer learning in fine-tuning large language models for domain-specific applications is illustrated. This study provides a robust framework for deploying intelligent auto-response systems in real-life. In the future, reinforcement learning techniques to mitigate current limitations will be explored, such as handling ambiguous prompts and improving response diversity.

Keywords: Transfer Learning, Domain-Specific Language Models, Fine Tuning, Generative Pre-trained Transformer, Automated Response

INTRODUCTION

Digital communication involving reliance on emails for collaborative information exchange and decision-making is fundamental in professional and research domains. However, the rapid increase in digital communication has also given rise to the challenge of email overload for employees and professionals. In sectors like energy sustainability and climate resilience, the sheer volume of emails can impede productivity and progress. According to Berners-Lee (2020), the environmental impact of digital communication, including email, is significant, with millions of emails exchanged daily. While individual email emissions may appear minimal, the aggregate effect contributes notably to the global carbon footprint (George, 2024). This highlights the broader

Cite as:

Osofuye, O.; Sogunle, O.; Osofuye A.; Ogunyale O, Jegede T.; Abiodun, T.; Oni. A. (2026). A Fine-Tuned GPT-2 For Intelligent Auto-Response System Using Transfer Learning with Contrived Email Data.*Journal of Science and Information Technology (JOSIT)*, Vol. 20, No. 1, pp. 1-14.

environmental concerns, such as climate change, which emphasises the necessity of finding more efficient, contextually relevant methods of email communication.

Automated email response systems have emerged as valuable tools for managing large volumes of email correspondence, reducing response times, and streamlining communication processes. However, the challenge lies in developing systems capable of generating contextually appropriate and coherent responses that meet the needs of diverse email interactions.

This study proposes an intelligent email auto-response system utilising contrived email transfer learning and Generative Pre-trained Transformer 2 (GPT-2) to address these challenges. Traditional email auto-response systems rely on rule-based or template-based approaches, which often produce generic and repetitive responses. Such systems are limited in their ability to adapt to complex and dynamic communication scenarios, leading to user frustration and diminished engagement (Arsovski et al., 2022).

Recent advancements in natural language processing (NLP) have introduced powerful language models such as GPT-2, which have demonstrated remarkable capabilities in generating human-like text (Chandrasekaran & Mago, 2021).

Despite the capabilities of pre-trained models like GPT-2 in natural language generation, their application in automated email responses is hindered by several challenges. Standard GPT-2 models are not optimised for specific domains such as email communication, leading to responses that may be contextually irrelevant. Recent work in domain adaptation for large language models like GPT-2 has shown that while such models perform well on general text, fine-tuning is necessary to ensure relevance and coherence in domain-specific tasks (Chronopoulou et al., 2021). Without fine-tuning, GPT-2 tends to generate responses that are either too vague or overly verbose, failing to meet the requirements of specific email interactions, such as concise and informative replies in professional settings. Also, email interactions can vary greatly, ranging from customer service queries to formal business communications and informal updates. A one-size-fits-all model struggles to adapt to these varying tones and contexts, reducing the effectiveness of the generated responses (Robertson et al., 2021).

To address these issues, this study aimed to develop an intelligent email auto-response system that can generate relevant and contextually appropriate responses across various email scenarios, by fine-tuning GPT-2 on a custom email dataset. This study seeks to evaluate the model's ability to generate coherent and contextually relevant email responses across various scenarios

GPT Models in Natural Language Processing

The Generative Pre-Trained Transformer (GPT) models have revolutionised the field of NLP by enabling the generation of coherent and contextually relevant text based on large-scale pre-training on diverse datasets. The original GPT model, introduced by OpenAI, demonstrated the potential of transformer architectures for various NLP tasks, including text generation, summarisation, and language translation (Minaee et al., 2020). Subsequent iterations, such as GPT-2 and GPT-3, expanded on this foundation by increasing model size and training on even larger corpora, thereby enhancing the ability to produce human-like text (Brown et al., 2020). GPT-2 has been widely adopted due to its balance between computational efficiency and generative capability. It utilises a multi-layer transformer network to predict the next word in a sequence, allowing it to generate contextually appropriate and fluent sentences.

Transfer Learning in NLP

Transfer learning has become a pivotal technique in NLP, enabling pre-trained language models to adapt to specific tasks with limited labelled data. This approach involves fine-tuning a model that has already learned general language patterns on a large corpus to a new, smaller dataset that represents the target domain. By retaining the knowledge acquired during pre-training, transfer learning allows models to achieve superior performance on specialised tasks with fewer resources compared to training from scratch (Zhuang et al., 2020).

Several studies have demonstrated the effectiveness of transfer learning in various NLP applications, such as sentiment analysis, machine translation, and question-answering. For instance, BERT (Bidirectional Encoder

Representations from Transformers) employs transfer learning to fine-tune on multiple downstream tasks, showcasing the versatility of pre-trained models (Koroteev, 2021). Similarly, this study employs transfer learning to adapt GPT-2 for the domain-specific task of email response generation, using a contrived email dataset to guide the model's learning.

Auto Response Systems: State of the Art

Automated response systems have been a focal point in both academic research and industrial applications, particularly for customer service and professional communication. Traditional systems often rely on predefined templates or rule-based frameworks, which limit their ability to handle complex and dynamic interactions.

While these systems can efficiently manage routine queries, they struggle with generating contextually appropriate responses for unexpected inputs. The advent of deep learning and NLP models, such as Seq2Seq and attention-based mechanisms, has improved the flexibility and coherence of auto-response systems (Huang, 2024). However, these models often require substantial training data and computational resources, making them challenging to implement in domains with limited data availability.

Recent efforts have focused on enhancing these systems through transfer learning and fine-tuning of pre-trained models, as demonstrated by studies using BERT and GPT-3 for dialogue generation (Labruna & Magnini, 2022) and customer support applications (Wang et al., 2021). There is also a novel deep learning technique in a smart email client that employed the Document2Vector (Doc2Vec) algorithm, and a hybrid model called GRU-Sent2Vec which combined the Gated Recurrent Unit (GRU) and Sentence to Vector (Sent2Vec) algorithms (Feng et al., 2020), where the authors aimed to develop a system that could generate sentence-level predictions or responses as compared to the word- level responses generated in earlier research. The Doc2Vec algorithm was evaluated to be better than the TF-IDF algorithm in predicting intelligent responses. However, the researchers lacked a large enough dataset to properly evaluate the GRU-Sent2Vec approach.

Arsovski et al. (2022) introduced a semi-automatic responder which differentiates between query and non-query emails and then generates automatic responses to the query emails. The query email dataset was divided into polar questions and factual questions. A polar question is a question whose answer can either be a "yes" or a "no", while a factual question is a question that can only have a fact as its answer and usually consists of question words (who, why, where, what, when etc.). The approach employed the use of Natural Language Processing, Neural Networks and Artificial Intelligence Markup Language.

Table 1 summarises the architecture for the intelligent email auto-response system which is proposed in this study. It involves a number of components with designated functions. Each element in the proposed architecture plays a distinct role in facilitating efficient communication and automatic response generation. Figure 1 serves to visualise the interactions of these elements, helping to better understand how the system functions as a whole.

Table 1. Summary of the architectural components of the proposed intelligent auto-response system.

Components	Functions
User Interface	Allows the user to read emails and send replies.
Email Transfer Mechanism	Handles the sending and retrieval of emails.
The Engine	Manages the processing of emails for sending, retrieval, and response generation.
Response Generator	Uses GPT-2 and transfer learning to generate intelligent responses.
Database	Stores recent emails for quick access by the user.
Email Server	The external server that emails are sent to and retrieved from.

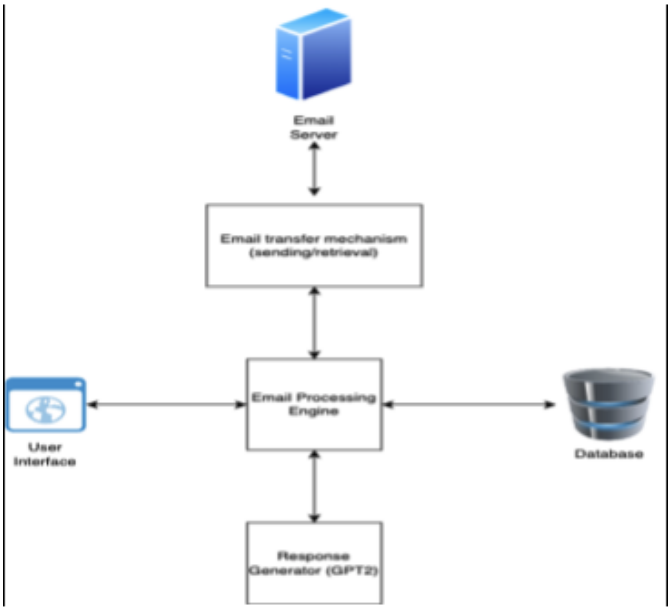


Figure 1. Architecture of the Proposed System.

METHODOLOGY

Dataset Description

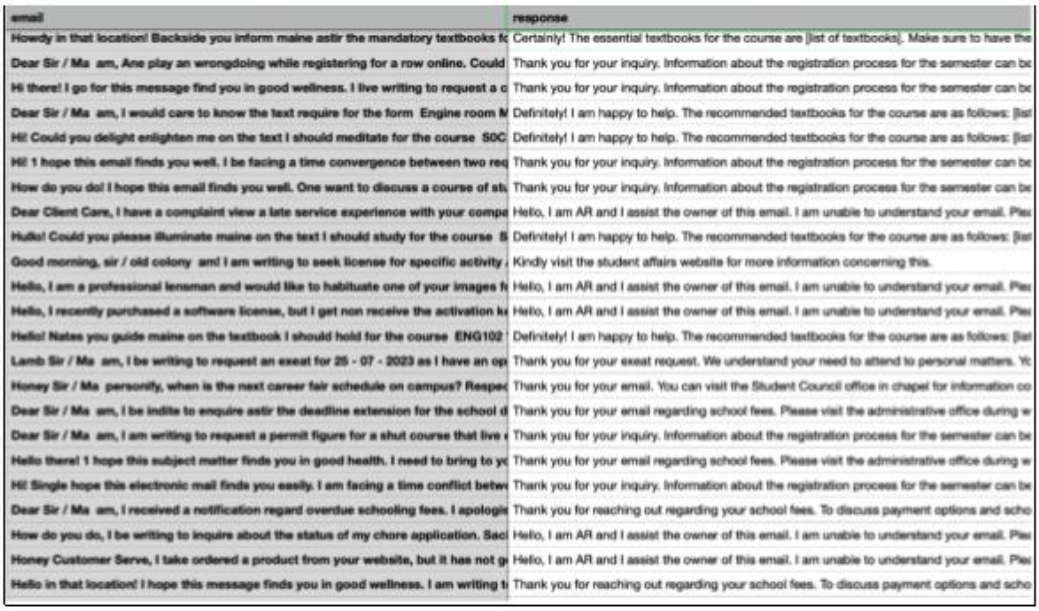
In this study, 2 datasets which were used in fine tuning include: the contrived email dataset and the Enron email dataset.

Contrived Email Dataset

A contrived dataset is an artificial dataset created to mimic real-world scenarios. Providing a controlled setting for training and evaluating machine learning models is a key advantage of this type of dataset. In email response systems, contrived datasets are useful in that they capture a wide range of communication styles and contexts, which may not be sufficiently represented in existing email datasets (Wescoat et al., 2022). They therefore can allow researchers to explore or investigate specific aspects of model behaviour. Some good examples of such aspects include handling ambiguous prompts and adapting to varying levels of formality. Contrived datasets have also been used in previous studies for carrying out tasks like dialogue generation, intent recognition, and even model training (Efiong et al., 2024).

In this study, a contrived dataset that mirrors the complexities of real-world email interactions was created with the hopes of enhancing the ability of the chosen model, GPT-2, to produce relevant responses across diverse email scenarios. The dataset represents a wide range of email scenarios, including customer service inquiries, informal exchanges mostly in a university setting, and so on. It has about 10,000 email-response pairs, each constructed to exhibit varying tones, levels of formality, and contextual complexity.

A combination of real-world examples and synthetic data augmentation techniques was used to generate email samples to ensure diversity and relevance. Figure 2 shows some rows of the created dataset. All rows, which individually consist of "email" and "response" columns, in the dataset were created to be contextually suitable and varied in terms of length and detail. The full dataset was then split into training (80%), validation (10%), and testing (10%) subsets.



email	response
Howdy in that location! Backside you inform maine astir the mandatory textbooks fo	Certainly! The essential textbooks for the course are [list of textbooks]. Make sure to have the
Dear Sir / Ma am, Ane play an wrongdoing while registering for a row online. Could	Thank you for your inquiry. Information about the registration process for the semester can be
Hi there! I go for this message find you in good wellness. I live writing to request a c	Thank you for your inquiry. Information about the registration process for the semester can be
Dear Sir / Ma am, I would care to know the text require for the form Engine room M	Definitely! I am happy to help. The recommended textbooks for the course are as follows: [list
Hi! Could you delight enlighten me on the text I should meditate for the course SOC	Definitely! I am happy to help. The recommended textbooks for the course are as follows: [list
Hi! I hope this email finds you well. I be facing a time convergence between two req	Thank you for your inquiry. Information about the registration process for the semester can be
How do you do! I hope this email finds you well. One want to discuss a course of stu	Thank you for your inquiry. Information about the registration process for the semester can be
Dear Client Care, I have a complaint view a late service experience with your compa	Hello, I am AR and I assist the owner of this email. I am unable to understand your email. Plea
Hullo! Could you please illuminate maine on the text I should study for the course S	Definitely! I am happy to help. The recommended textbooks for the course are as follows: [list
Good morning, sir / old colony and I am writing to seek license for specific activity .	Kindly visit the student affairs website for more information concerning this.
Hello, I am a professional lensman and would like to habituate one of your images fo	Hello, I am AR and I assist the owner of this email. I am unable to understand your email. Plea
Hello, I recently purchased a software license, but I get non receive the activation ke	Hello, I am AR and I assist the owner of this email. I am unable to understand your email. Plea
Hello! Notes you guide maine on the textbook I should hold for the course ENG102	Definitely! I am happy to help. The recommended textbooks for the course are as follows: [list
Lamb Sir / Ma am, I be writing to request an exeat for 25 - 07 - 2023 as I have an op	Thank you for your exeat request. We understand your need to attend to personal mathers. Yo
Honey Sir / Ma personify, when is the next career fair schedule on campus? Respec	Thank you for your email. You can visit the Student Council office in chapel for information co
Dear Sir / Ma am, I be indite to enquire astir the deadline extension for the school d	Thank you for your email regarding school fees. Please visit the administrative office during w
Dear Sir / Ma am, I am writing to request a permit figure for a shut course that live	Thank you for your inquiry. Information about the registration process for the semester can be
Hello there! I hope this subject matter finds you in good health. I need to bring to yo	Thank you for your email regarding school fees. Please visit the administrative office during w
Hi! Single hope this electronic mail finds you easily. I am facing a time conflict betwe	Thank you for your inquiry. Information about the registration process for the semester can be
Dear Sir / Ma am, I received a notification regard overdue schooling fees. I apologis	Thank you for reaching out regarding your school fees. To discuss payment options and scho
How do you do, I be writing to inquire about the status of my chore application. Sac	Hello, I am AR and I assist the owner of this email. I am unable to understand your email. Plea
Honey Customer Serve, I take ordered a product from your website, but it has not g	Hello, I am AR and I assist the owner of this email. I am unable to understand your email. Plea
Hello in that location! I hope this message finds you in good wellness. I am writing t	Thank you for reaching out regarding your school fees. To discuss payment options and scho

Figure 2. A sample of the rows and columns of the contrived dataset.

The Enron Email Dataset

The original dataset of Enron emails comprises around 500,000 emails that were authored by Enron Corporation employees. This dataset was acquired by the United States Federal Energy Regulatory Commission as part of their investigation into the company's collapse. The processed version utilised in this study was sourced from Feng et al. (2020) and contains approximately 19871 email pairs arranged in rows, featuring two columns: the "received" column and the "response" column. It should be noted that further cleaning and pre-processing were carried out. Figure 3 shows the first 20 rows of the dataset.

Preprocessing Techniques

Effective preprocessing is crucial for preparing the contrived email dataset for training the GPT-2 model. The following preprocessing steps were applied. Each email and response was tokenised using the Byte Pair Encoding (BPE) technique, which is the standard for GPT-2 models to ensure that even out-of-vocabulary words were handled efficiently. This process converts text into a sequence of tokens, representing words or subwords, which the model can process.

To maintain consistency and reduce noise, the text was first normalised. This step involved converting all the characters to lowercase, also removing special characters, handling of contractions and so on. To help the model recognise structure and improve context-awareness, some special tokens such as, etc. were then introduced to mark different sections of the email data. Emails and responses were padded or truncated to a fixed length of 512 tokens to ensure uniformity in model input.

Texts that were longer than necessary were truncated to work with this limit while shorter texts were padded with a special token. For the sake of robustness, data augmentation techniques like paraphrasing, synonym replacement etc. were also applied to the training set. This was to increase the variability of the input data without change in the original and intended meanings.

Transfer Learning (with GPT-2)

As a result of GPT-2's strong performance in natural language generation tasks, as well as its capacity to learn complex patterns from large-scale text data, it was chosen for this study. Fine-tuning on the contrived email dataset then followed.

GPT-2 is based on the transformer architecture, which consists of multiple layers of self-attention and feed-forward neural networks. The model’s core component of Multi-Head Attention Mechanism allows the model to attend to different parts of the input sequence simultaneously, capturing various aspects of context and dependencies between tokens.

The GPT-2 model used in this study has 12 layers, 12 attention heads, and 117 million parameters, providing a good balance between computational efficiency and performance. Figure 4 shows the architecture of the Generative Pre-Trained Transformer upon which GPT-2 was built (Wolfe, 2022).

GPT-2 was selected as the base model for this study due to its open-source availability via the Hugging Face Transformers library, which ensures experimental reproducibility and allows for fine-grained control over the transfer learning process. Unlike larger, closed-source models accessible only via API, GPT-2 provides a computationally efficient framework suitable for domain-specific fine-tuning on specialised datasets while maintaining high performance in text synthesis.

enron-feng

	Received	Response
0	resumes of whom ?	The commercial support people that you and Hunter want to make commercial managers .
1	Christy I read these points and they definitely need some touch up . I don t understand why we need to	Phillip To the extent that we can give Chair Hoecker our spin on the reasons for the hikes we w
2	Phillip I have a meeting tomorrow morning with the Oregon PUC staff to discuss a number of pricing and	Yes you can use this chart . Does it make sense ?
3	Phillip Which one should I do the x is half the price . Jeff	I like the cedar t version better . Why don t you tell this shop teacher not to be to comfortable
4	Phillip The Social Security and Medicare tax has been previously withheld on your deferral . You would b	Thanks
5	Phillip This is the candidate I spoke with you about this morning . . . Brent currently works in our London	Adrianne I cannot download his resume . Please resend . Phillip
6	Phillip Lets try this one . . . .Thanks ! Adrianne	Left message and sent email . I will let you know as soon as I have spoken to him .
7	Phillip I need to get the contract for Galaxy and Grande executed today . In addition I will need to close t	Greg I would rather sign and fax copies of the necessary contracts then follow up with original
8	Phillip I am waiting to get info . on two more properties in San Antonio . The broker will be faxing info . o	Jeff Can you resend the info on the three properties you mailed and the one you faxed on Tues
9	Bernie Good Morning . I hope all is well . I have tried not bother you with this project during Q but now I	Kirk I ve added my comments . See attached Given Enron s current situation can you give a st
10	For purposes of an accelerated distribution from the PSA a single sum distribution in Section . means th	David For clarity wherever you cite Section . in your response I assume you mean Section . . Y
11	Phillip there are a number of alternative systems that will allow the same level of energy efficiency . I wou	Richard I checked with Wink and his prices were slightly lower per square foot but the builder
12	Phillip Could you please do me a favor ? I would like to read your current title policy to see what it says	Bob I found the warranty deed but I could not find the title policy . I will call the title company t
13	Phillip Pursuant to your request please see the attached . Thanks Renee	Renee Thank you for the forms . I am trying to estimate my tax obligation . It is my understandi
14	Phillip Keith Attached is the first draw request I will need some of these funds immediately . I think check	Greg I would rather sign and fax copies of the necessary contracts then follow up with original
15	Sheri and I would like to discuss the practice questions and graphic ideas with you for the the Knowledg	Can we meet from am am instead ?
16	Phillip This section pertains to terminated employees who are paid out in the year following the terminati	Renee I would like to have a meeting to go over the payment methodology . Scott Neal and To
17	Heather This is exactly what we need . Would it possible to add the prior day for each of the dates below	Please let me know if you still need Curve Shift . Thanks Heather
18	David For clarity wherever you cite Section . in your response I assume you mean Section . . Your respor	Pam Can you help with below ? David
19	Heather Did you attach the file to this email ?	Please let me know if you still need Curve Shift . Thanks Heather
20	Renee Thank you for the forms . I am trying to estimate my tax obligation . It is my understanding that th	Phillip The Social Security and Medicare tax has been previously withheld on your deferral . Yo

Figure 3. A sample of the rows and columns of the pre-processed Enron dataset.

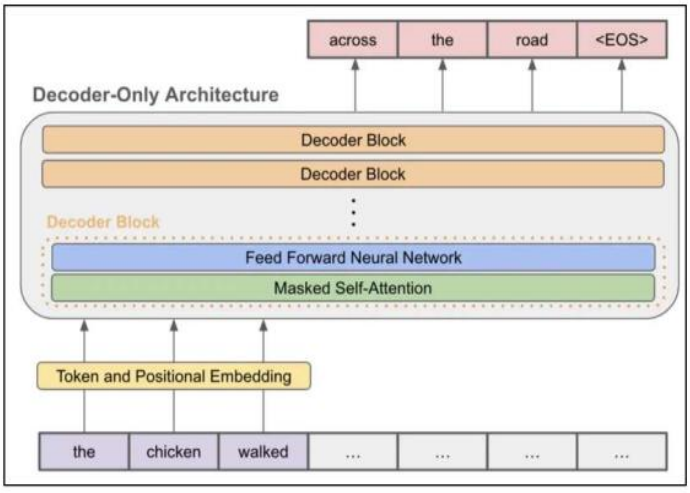


Figure 4. GPT Architecture.

Fine-tuning Process on the dataset

Fine-tuning involved re-training GPT-2 on the smaller contrived email dataset for email response generation. To start with, email response pairs, already pre-processed, were formatted into input sequences suitable for GPT-2. This meant every single sequence would have an email prompt followed by the corresponding response, separated by special tokens.

After extensive hyperparameter tuning, a learning rate of 3e-5, a batch size of 16, and a maximum sequence length of 512 tokens were selected. In addition, the Adam optimiser was applied to minimise the loss function. In a bid to prevent overfitting, early stopping was used, and model checkpoints were saved at regular intervals. The best-performing model, based on validation loss,

was then selected for final evaluation. Other techniques like dropout, weight decay etc. were applied to the model during fine-tuning to improve generalisation and reduce overfitting.

Evaluation Metrics

The 2 evaluation metrics used are the BLEU and ROUGE score.

1. **BLEU (Bilingual Evaluation Understudy) Score:** This was used to check how similar the generated responses are to the test set's reference responses while helping to measure response fluency and relevance.
2. **Recall-Oriented Understudy for Gisting Evaluation (ROUGE) Score:** This was used to check the quality of responses by comparing the overlap of "n-grams", "word sequences," and "word pairs" between the generated and reference responses.

Human evaluators also rated the generated responses on criteria such as relevance, coherence, appropriateness etc. (qualitative assessment) to identify strengths and weaknesses not captured by automated metrics.

Experimental Setup

The base GPT-2 was obtained from the Hugging Face Transformers library. The model obtained was initialised with weights from the library as well. The training environment consisted of the tools and configurations listed in Table 2. Weights and Biases helped to track model performance in real-time.

Experimental Scenarios and Configurations

To evaluate the performance of the fine-tuned GPT-2 model, three experimental scenarios were set up as detailed in Table

Table 2. Model training environment.

	Description
Hardware	NVIDIA A100 GPU with 40 GB memory was utilised for efficient parallelised training. CPU configurations included 32-core Intel Xeon processors.
Software	Python 3.9, Hugging Face's Transformers, PyTorch, Pandas, Scikit-learn.
Development Environment	Google Colab Pro/Kaggle, Git.

Table 3. Experimental scenarios and configurations

Model	**Experiment**
Base GPT-2 tested against unseen Enron emails	These served as the baselines, allowing us to measure how the raw, unmodified GPT-2 performed in generating intelligent auto-responses.
Base GPT-2 tested against unseen contrived emails	
GPT-2 fine-tuned with Enron emails tested against unseen Enron emails	Trained for 10 epochs, batch size 16, learning rate 3e-5, and early stopping at 3 epochs to prevent overfitting.
GPT-2 fine-tuned with contrived emails tested against unseen contrived emails	The fourth scenario involved fine-tuning GPT-2 on the contrived email dataset using transfer learning. This setup tested the model's ability to learn specificities of email exchanges.

	In this scenario, the email dataset was pre-categorised into different contexts (e.g., customer service, professional inquiries, and informal communication). The model was fine-tuned to allow it to understand the type of email and generate more context-aware responses. This variant was evaluated to determine whether adding contextual information improved the relevance and appropriateness of the responses.

RESULTS AND ANALYSIS

Dataset Description

The BLEU score was used for quantitative analysis while the ROUGE score was used for the comparative analysis.

Quantitative Evaluation

In this study, the BLEU score was calculated for 4 instances of the fine-tuned GPT-2 model. The instances and corresponding BLEU score are displayed in Table 4.

Bi-Lingual Evaluation Understudy Score (BLEU Score)

For this study, the BLEU score was calculated for 4 instances of the fine-tuned GPT-2 model. The instances and corresponding BLEU scores are displayed in the table. Table 5 shows the BLEU scores obtained from evaluating different instances of GPT-2.

Table 4. BLEU Scores of GPT-2 Instances.

Instance	BLEU Score
GPT-2 fine-tuned with Enron emails tested against unseen Enron emails	0.66
GPT-2 fine-tuned with contrived emails tested against unseen contrived emails	0.60
Base GPT-2 tested against unseen Enron emails	0.54
Base GPT-2 tested against unseen contrived emails	0.53

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) Score

Three types of ROUGE scores namely the ROUGE-1, ROUGE-2 and ROUGE-L are calculated for the four instances of the GPT-2 model. Each metric has three scores: 'r' (recall), 'p' (precision), and 'f' (F1 score).

ROUGE-1 measures the measures the overlap of unigram (single-word) sequences between the generated text and the reference text. ROUGE-2 measures the overlap of bigram (two-word) sequences while ROUGE-L measures the longest common subsequence (LCS) between the generated and reference summaries, considering both unigrams and longer sequence matches. The ROUGE scores for the four instances of the model are shown in Tables 5, 6, 7, and 8.

Table 5. ROUGE scores for GPT-2 re-trained and tested with Enron emails.

ROUGE Variant	Recall	Precision	F1-Score
ROUGE-1	0.128	0.120	0.103
ROUGE-2	0.009	0.009	0.008
ROUGE-3	0.091	0.090	0.074

Table 6. ROUGE scores for GPT-2 re-trained and tested with contrived emails.

ROUGE Variant	Recall	Precision	F1-Score
ROUGE-1	0.625	0.555	0.580
ROUGE-2	0.437	0.422	0.425
ROUGE-3	0.612	0.545	0.569

Table 7. ROUGE scores for Base GPT-2 tested against sample Enron emails.

ROUGE Variant	Recall	Precision	F1-Score
ROUGE-1	0.212	0.035	0.057
ROUGE-2	0.002	0.0003	0.006
ROUGE-3	0.163	0.025	0.041

Table 8. ROUGE scores for Base GPT-2 tested against sample contrived emails.

ROUGE Variant	Recall	Precision	F1-Score
ROUGE-1	0.127	0.079	0.192
ROUGE-2	0.004	0.004	0.004
ROUGE-3	0.108	0.070	0.080

A plot of precision values against recall values of the ROUGE instances of 2 models is shown in Figure 5. The first model with lower values is GPT-2 re-trained and tested with Enron emails, while the second with higher values is GPT-2 re-trained and tested with contrived emails.

Additionally, human evaluation was applied on a subset of 200 email responses to rate the quality based on fluency, relevance, and appropriateness. On a scale of 1 to 5, the fine-tuned model achieved an average score of 4.1, compared to 3.2 for the baseline model.

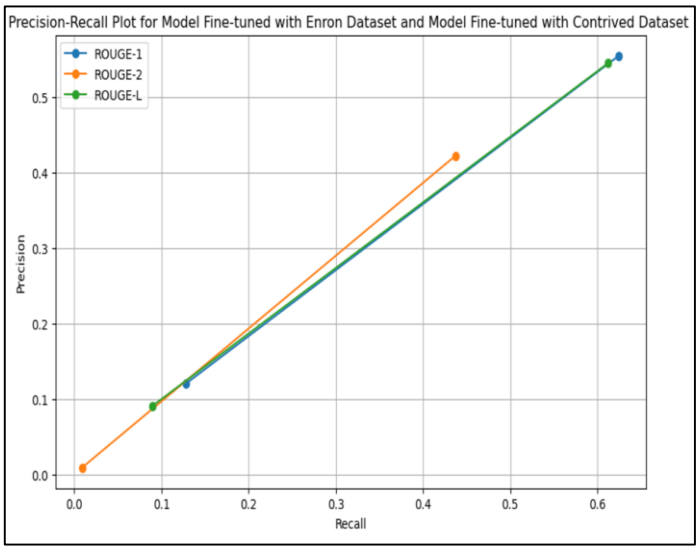


Figure 5. Precision-Recall Plot of the GPT-2 models fine-tuned with the Enron and Contrived datasets.

Comparative Analysis with Baseline Models

A comparative analysis between the baseline GPT-2 and the fine-tuned GPT-2 across the three experimental scenarios demonstrated the effectiveness of the transfer learning process:

1. **Baseline GPT-2 Performance:** The unmodified GPT-2 struggled with generating contextually appropriate email responses, as reflected in its low ROUGE scores. For ROUGE-1, the recall was 0.127 with an F1 score of 0.192, indicating poor alignment between generated and reference responses. ROUGE-2 and ROUGE-3 scores were even lower, with F1 values of 0.004 and 0.080 respectively, demonstrating its inability to capture meaningful multi-word sequences or deeper context.
2. **Enron Fine-Tuned GPT-2***:* Higher ROUGE-1 recall and F1 values of 0.128 and 0.103, but with only marginal improvements in ROUGE-2 and ROUGE-3 scores (F1 scores of 0.008 and 0.074). Even though the model had learnt how to replicate some patterns from the Enron dataset, its ability to generalise across diverse communication scenarios was still quite limited.
3. **Contrived Fine-Tuned GPT-2:** ROUGE-1 scores increased substantially, with recall at 0.625 and F1 at 0.580. Also, ROUGE-2 and ROUGE-3 F1 scores rose to 0.425 and 0.569, respectively. These values show the effectiveness of the contrived dataset in training the model to generate coherent and contextually relevant responses.

This comparative analysis shows that when GPT-2 is fine-tuned on a well-structured task-specific dataset, in this case, the contrived dataset, it can greatly enhance the model's capacity to give better and more meaningful responses.

DISCUSSION

Error Analysis

While the fine-tuned GPT-2 showed notable improvements, certain errors persisted as shown in Table 9. Addressing these issues would require further experimentation with reinforcement learning techniques or dataset refinement to better handle complex conversation dynamics.

Impact of Transfer Learning on GPT-2 Performance

By fine-tuning the GPT-2 model on a domain-specific contrived email dataset, the general-purpose language model was successfully adapted to the specialised task of email response generation and exhibited improved comprehension of email context, including the intent of the sender and the appropriate tone for the response. This adaptation was evident in the model's ability to generate responses that aligned closely with the expected communication style. Fine-tuning on the contrived email dataset helped reduce generalisation errors common in pre-trained models. The model became less likely to generate overly generic or irrelevant responses, instead producing outputs that were more tailored to the specific content of the prompt emails.

The use of transfer learning led to notable improvements in quantitative metrics such as BLEU, and ROUGE scores. This indicates that the model learned to generate responses that were not only fluent and coherent but also closely matched the reference responses. From the BLEU scores, it can be deduced that the Enron fine-tuned and contrived fine-tuned models had similar performances. While the BLEU score is a good metric, the ROUGE score can also provide further insight into the quality of text generated by the model.

Based on Tables 5, 6, 7, and 8, it can be deduced that GPT-2 re-trained with the contrived dataset seemed to have the best performance out of all the instances, especially better than the GPT-2 instance re-trained with Enron emails. This further supports the idea that structured unlabelled data will tend to provide better performance when used for fine-tuning instead of unstructured and generalised unlabelled data. Figure 6 shows some responses generated by the GPT-2 model fine-tuned with the contrived dataset when tested with an unseen portion of the contrived dataset.

Table 9. Description of errors from the fine-tuned model.

Error	Description
Repetitiveness	In some cases, the model produced repetitive phrases when the input prompt was vague or ambiguous. For instance, a general question like "Can you help me with this issue?" sometimes led to circular responses such as "Yes, I can help. How can I assist with this issue?"
Overly Polite or Neutral Responses	In some instances, the model defaulted to very neutral or overly polite responses, especially in customer service contexts. While politeness is generally positive, some scenarios required a more direct tone that the model failed to provide.

It should be noted that the data augmentation techniques used to create the contrived dataset are a major factor affecting the quality of the dataset as some sentences are not as high quality as they should be.

Insights on Intelligent Auto Response Systems

How auto-response systems perform is majorly based on the quality and relevance of their training data. When there are contrived datasets that accurately represent the target domain, it is possible to significantly improve the models' performances, as these datasets can help to provide the necessary context and nuances.

Automated systems are great for handling large volumes of routine queries, however, striking the right balance between automation and personalisation is key. This study demonstrates that fine-tuning large language models like GPT-2 on domain-specific data can bring forth more personalised responses, creating better user experiences and can be quite powerful for many applications like supporting customers, virtual assistants, etc.

RECOMMENDATIONS

Addressing key challenges in the development of intelligent systems needs a collaborative effort. As a result, this study helps provide the following actionable recommendations that can guide future initiatives that align with ethical AI practices, sustainable computing, and even advancements in natural language processing:

1. Government and policy stakeholders should encourage the development of open datasets for NLP research, particularly in underrepresented domains such as workplace communication, while ensuring strict data privacy and ethical guidelines.

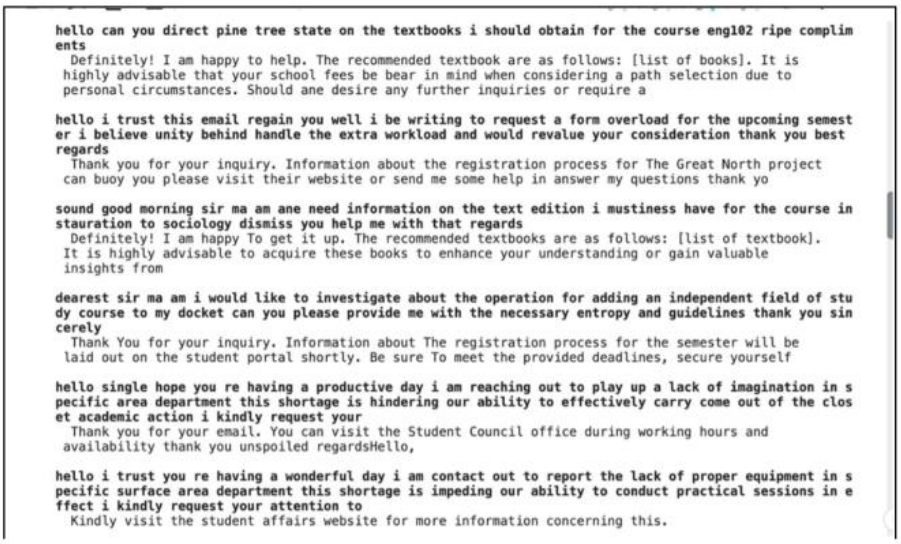


Figure 6. Responses generated by the GPT-2 model fine-tuned with the contrived dataset.

2. Energy institutions should support AI applications, especially those that reduce computational energy requirements for training large language models, thereby contributing to sustainable AI practices.
3. Researchers should prioritise diverse datasets that reflect real-world variability, and explore advancements like reinforcement learning and larger transformer models to enhance response quality.

Limitations and Future Directions

It would be quite important to work on expanding the contrived dataset (although it was useful for initial testing and evaluation) with more varied and authentic email exchanges for a more generalisable and robust model. This is because it does not fully capture how complex real-world email interactions can be.

Training and fine-tuning large models like GPT-2 require significant computational resources. This constraint can limit the ability to experiment with larger models or extensive hyperparameter tuning, especially in resource-limited settings. The smaller model sometimes struggled with ambiguous or poorly defined email prompts, leading to less accurate or relevant responses.

Data privacy and ethical use also need to be ensured while implementing intelligent auto-response systems in the real-world (ethical deployment of AI). It is important that stakeholders ensure that user data is handled responsibly, and that responses generated should not propagate biases.

It would also be good if future works could focus on incorporating more authentic email exchanges, exploring more advanced transformer models and fine-tuning with reinforcement learning techniques. Integrating with existing email platforms would also be a great next step.

CONCLUSION

The contrived fine-tuned GPT-2 model that was developed in this study demonstrated significant improvements over baseline models in generating contextually appropriate, fluent, and relevant email responses. The used evaluation metrics, including ROUGE scores, helped validate the effectiveness of transfer learning and the importance of using domain-specific data for training. This helps to show the value of contrived email datasets in the initial stages of developing and evaluating email response systems. The key takeaways from this study include:

1. Although there are limits to how generalisable they are, contrived datasets can help to effectively bootstrap training whenever real-world data is unavailable.

2. Fine-tuning on task-specific datasets significantly improves language models for specialised applications.

3. Ambiguities in email prompts revealed areas for further improvement in handling incomplete or vague inputs.

Email overload is known to contribute to employee stress and decreased productivity (Lanctot & Duxbury, 2021). This research addresses these challenges by improving the generation of automated email responses, streamlining communication and reducing cognitive load for professionals. The fine-tuned model in this study produced high-quality, context-aware responses that were both fluent and relevant, outperforming baseline models on several evaluation metrics.

Authors' Contributions

Odunayo Damilola OSOFUYE: Supervision, Conceptualization, Validation, Formal analysis, Methodology, Writing - Original draft preparation

Olubusolami SOGUNLE: Data curation, Software, Methodology, Visualization, Investigation, Writing – Original draft preparation

Ariyo Adesegun, OSOFUYE: Software, Visualization, Investigation, Writing - Reviewing and Editing.

Oluwaseun Ayokunle, OGUNYALE: Software, Visualization, Investigation, Writing - Reviewing and Editing

Tope John, JEGEDE: Software, Visualization, Investigation, Writing - Reviewing and Editing

Theresa Nkechi, ABIODUN: Software, Visualization, Investigation, Writing - Reviewing and Editing

Aderonke Atinuke, ONI: Supervision, Project administration, Methodology, Writing - Original draft preparation.

All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

Data will be made available by the corresponding author on request.

ACKNOWLEDGEMENT

The publication support provided by the Covenant University Centre for Research, Innovation and Discovery is gratefully acknowledged.

Conflicts of Interest

The authors declare no conflicts of interest.

REFERENCES

Arsovski, S., Oladele, M., Cheok, A., Premcevski, V., & Markoski, B. (2022). An approach to email categorization and response generation. *Computer Science and Information Systems*, *19*(2), 913–934. https://doi.org/10.2298/csis211101009a

Berners-Lee, M. (2020). *How bad are bananas? The carbon footprint of everything.* Profile Books Ltd.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv (Cornell University)*, *33*, 1877–1901. https://doi.org/10.48550/arxiv.2005.14165

Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity—A survey. *ACM Computing Surveys*, *54*(2), 1–37. https://doi.org/10.1145/3440755

Chronopoulou, A., Peters, M. E., & Dodge, J. (2021). Efficient hierarchical domain adaptation for pretrained language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2112.08786

Efiong, J. E., Akinsola, J. E. T., Akinyemi, B. O., Olajubu, E. A., & Aderounmu, G. A. (2024). A contrived dataset of substation automation for cybersecurity research in the smart grid networks based on IEC61850. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *22*(5), 1320. https://doi.org/10.12928/telkomnika.v22i5.26000

Feng, Y., Naeem, M. A., Mirza, F., & Tahir, A. (2020). Reply using Past Replies—A deep Learning-Based E-Mail client. *Electronics*, *9*(9), 1353. https://doi.org/10.3390/electronics9091353

George, D. (2024). Digital hoarding: the rising environmental and personal costs of information overload. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.12802575

Koroteev, M., V. (2021, March 22). *BERT: A review of Applications in Natural Language Processing and Understanding*. arXiv.org. https://arxiv.org/abs/2103.11943

Labruna, T., & Magnini, B. (2022). Fine-Tuning BERT for Generative Dialogue Domain Adaptation. In *Lecture notes in computer science* (pp. 513–524). https://doi.org/10.1007/978-3-031-16270-1_42

Lanctot, A., & Duxbury, L. (2021). When everything is urgent! Mail use and employee well-being. *Computers in Human Behavior Reports*, *4*, 100152.

https://doi.org/10.1016/j.chbr.2021.100152

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2020). Deep Learning based text Classification: A Comprehensive review. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2004.03705

Robertson, R. E., Olteanu, A., Diaz, F., Shokouhi, M., & Bailey, P. (2021). "I can't reply with that": Characterizing problematic email reply suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 724, pp. 1–18). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445557

Wang, P., Fang, J., & Reinspach, J. (2021, November). CS-BERT: A pretrained model for customer service dialogues. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI* (pp. 130–142). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nlp4convai-1.13

Wescoat, E., Kerner, S., & Mears, L. (2022). A comparative study of different algorithms using contrived failure data to detect robot anomalies. *Procedia Computer Science, 200, 669–678.* https://doi.org/10.1016/j.procs.2022.01.265

Wolfe, C. R. (2022, June 6). *Understanding the Open Pre-Trained Transformers (OPT) library*. Towards Data Science. https://medium.com/towards-data-science/understanding-the-open-pre-trained-transformers-opt-library-193a29c14a15

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43–76. https://doi.org/10.1109/jproc.2020.3004555