# Fine-Tuning Large Language Models for Domain-Specific Response Generation:A Case Study on Enhancing Peer Learning in Human Resource

MSc Research Project
Data Analytics

Diksha Bhatnagar
Student ID: 21183961

School of Computing
National College of Ireland

Supervisor: Professor Rejwanul Haque

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Diksha Bhatnagar |
| **Student ID:** | 21183961 |
| **Programme:** | Data Analytics |
| **Year:** | 2022-23 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Professor Rejwanul Haque |
| **Submission Due Date:** | 16/08/2023 |
| **Project Title:** | Fine-Tuning Large Language Models for Domain-Specific Response Generation:A Case Study on Enhancing Peer Learning in Human Resource |
| **Word Count:** | 2609 |
| **Page Count:** | 20 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Diksha Bhatnagar |
| **Date:** | 30/08/2023 |

## PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ☐ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ☐ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ☐ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Fine-Tuning Large Language Models for Domain-Specific Response Generation:A Case Study on Enhancing Peer Learning in Human Resource

Diksha Bhatnagar
21183961

**Abstract**

This research delves into front in-line Natural Language Processing (NLP) techniques, focused on devising innovative solutions specifically tailored for small-scale organisations to enhance precision and efficiency in the Human Resources domain. Keeping an employee-centric view in mind, the study molds Large Language Models (LLM) to excel in this domain.

The architecture focuses on generating contextually relevant answer prompts directed towards overall employee development, peer learning, and corporate culture. The strategy is underpinned by the utilisation of employee survey data on which the model is trained to glean insights from anonymised and consent-obtained responses. This synergy between natural language processing and survey data combines to fuel the system to offer accurate and contextually aware answers. Additionally, the dissertation explores the novel concepts of text synthesis, treating it as a self-contained entity. This intriguing avenue explores as well as promises potential applications in communication enhancements, though its inner workings are considered enigmatic, akin to a black box.

This study places a high priority on ethical concerns and compliance, ensuring the appropriate use of employee data and adherence to ethical research practices. The dissertation discusses the difficulties with ethics and compliance that were encountered during the study process and suggests solutions. It examines the significance of open data usage, clear consent, and the implementation of a finely tuned language model.

Integration of NLP with employee-centric data and venturing into the zone of multimedia synthesis, this study contributes to the rapidly growing field of AI-driven corporate solutions. The results demonstrate the effectiveness of specialised NLP methods in increasing communication dynamics and serve as a foundation for further study in this broad area.

## 1 Introduction

In today's rapidly evolving corporate landscape, effective information exchange and seamless communication are of great importance, serving as cornerstones of success. Among the cutting-edge technologies and automations, playing a significant role in driving the transformation in Natural Language Processing (NLP) is the desire of many small-scale organizations. NLP serves as an interdisciplinary field at the intersection of computer

science and linguistics. The advent of NLP is expected to reshape the way organisations engage with their employees and customers, granting machines the capability to comprehend, interpret, and generate human language.
In recent times, the deployment of NLP has moved beyond traditional paradigms, finding a welcoming environment in corporate settings and thriving in employee-centric communications. This dynamic environment underscores the need to establish efficient communication channels between businesses and their staff. Employees, who form the backbone of every business, require not only prompt solutions to their inquiries but also personalized interactions that are relevant to the situation. NLP emerges as a powerful tool for meeting these demands and bridging the technological and interpersonal gaps.
Information retrieval and effective communication play a significant role in environment-centric contexts. Therefore, an organisation's ability to respond quickly to employee inquiries, foster cooperation, and efficiently disseminate knowledge is crucial. The timely provision of accurate information can have a substantial impact on employee satisfaction, engagement, and productivity, whether it involves defining corporate policies, addressing HR-related matters, or enhancing the on-boarding process.

The increasing use of NLP techniques has led to a compelling shift in the corporate communication paradigm and how these organisations are redesigning their ecosystems. For instance, the ability to promptly address employee inquiries, disseminate knowledge, etc., which in turn significantly impacts employee satisfaction, engagement, and productivity. NLP emerges as a crucially in comprehending, summarising, and extracting insights from this deluge of information, as organisations grapple with the challenges of managing vast volumes of textual data. The potential of NLP techniques to enhance communication extends beyond just text it also encompasses the realm of audio synthesis, ushering in a new era of engaging and interactive workplace interactions.
In the context of large language models, transfer learning is the process of pre-training a model on a substantial dataset and then fine-tuning it for a specific task or domain. This approach has completely revolutionised the field of Natural Language Processing (NLP) and has led to the development of several highly potent language models, such as Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT), and Text-to-Text Transfer Transformer (T5).

With these future developments in mind, this dissertation concentrates on leveraging the potential of NLP in corporate settings and delivering a tailored solution. This solution not only offers employees precise answers to their queries but also enhances organisations' overall communication infrastructure through the utilisation of advanced fine-tuning techniques and the wealth of employee survey data. Furthermore, this work extends the boundaries of conventional communication methods by exploring audio synthesis.
Many research studies have focused on a model-driven approach to defining research questions. However, this study makes an attempt to categorise respondents based on their responses. It seeks to establish a broader relationship using a representative survey sample, which typically reveals limited information about individual preferences (Tibor et al.; 2023). Here, we will train and fine-tune multiple Large Language Models on the survey response dataset, and then we will identify the most efficient model to consider when working in a resource-constrained environment with minimal incurred costs.

The high level architecture of the project is depicted in Figure 1

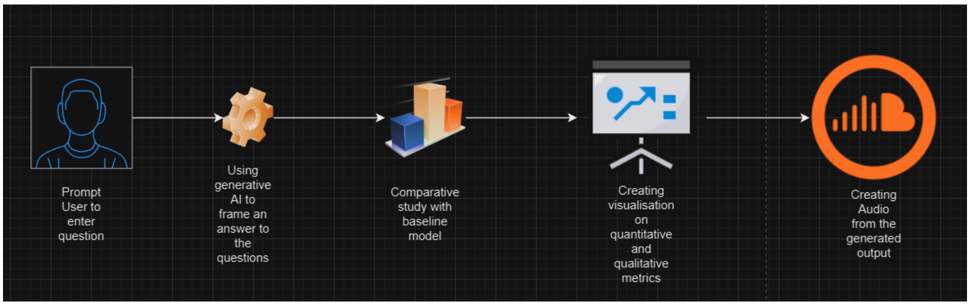



Figure 1: Project Architecture (High Level)

## 2 Related Work

The development of large-scale pre-trained language models like Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer 2 (GPT-2), and Text-to-Text Transfer Transformer (T5), have recently transformed the fields of natural language processing (NLP) and machine learning. In numerous NLP tasks, such as text generation, translation, and sentiment analysis, these models have proven to have excellent ability. The efficiency of these models in producing answers to questions with a specified domain has been investigated in earlier studies. Attention has also been drawn to the use of transfer learning approaches to hone these models for particular purposes, such as producing responses in the context of employee survey data.

### 2.1 Expensive LLM's and Pre-Conceptions

In one of the study authors suggested LLMs might develop prejudice, reinforce pre-conceptions in the training dataset, and convey incorrect information as true. Here a review on benefits and issues related to AI were discussed. Where Google referred to Bard and other LLMs like BLOOM as an open access, multilingual LLM with 176B parameters that can be used to deploy additional programmes and power chat-like user interfaces. Stating capacity to reliably induce desired outcomes is not always simple, especially given the sensitivity to the precise text prompt choice as suggested by (Moslem et al.; 2023).

LLM-based chatbots face a fundamental problem in that they have the potential to "hallucinate," or offer erroneous information as true. Large Language Models (LLMs), which have been used for some time and have varied degrees of ability to produce meaningful and valuable text, are not unique to ChatGPT.[1] However, a lot of these technologies are openly accessible as command line interfaces (CLIs), which may account for ChatGPT's rapid climb to renown. The GPT-3 model was skillfully developed by OpenAI to be implemented at scale alongside an easy-to-use user interface.The computer hardware required to develop AI models is expensive to produce and uses a lot of energy both during development and deployment.Without taking into account research, development, and energy use, the cost of the hardware is estimated to be close to 5 million dollars based on the processing power that OpenAI has access to (10,000 Nvidia V100s).

[1] `https://openai.com/`.

According to a recent analysis, training BERT (a 6 billion parameter LLM) to completion would generate between 21 and 78 metric tonnes of CO2. Recall that ChatGPT is a 175 billion parameter LLM and that this only takes into account the cost of training, not any cost associated with producing text in response to possibly billions of daily questions (G et al.; 2023). Hence in a resource constrained situation and keeping environmental constraints in mind its important to understand to efficiently use the existing resources instead of creating other models at large scale.

## 2.2 Transfer Learning and Optimisation

Yu et al.2023 suggests transfer learning using Pre-trained Language Models (PLMs) has become a potent approach to address issue. PLMs are strongly dependent on large, labelled datasets, which limits their ability to quickly adjust to changing viewpoints. To take advantage of pre-trained information and get around dataset restrictions, techniques like hyper-parameter tuning for prompt generation arose. However, these methods are still insufficient to guarantee accuracy or to produces a single shot or zero shots scenarios. Another study suggests Large language models have revolutionised automatic text summarisation, but a challenge for this work continues to be the lack of in-domain datasets.Hence synthetic and real in-domain data are used in tandem through an iterative data augmentation strategy, greatly enhancing the summarisation procedure creating a "Transformer" language-based abstract text summariser for Arabic content. The effectiveness of the proposed model is assessed by optimising the BART model (Jezia et al.; 2022). Here the study focused on transfer learning paradigm of machine learning to train models using synthetic as well as real world data.

## 2.3 Biases

One more study focused on Large-scale language models' reliance on copious amounts of textual training data, drawn from a variety of text sources like Wikipedia, BookCorpus, and the enormous web, has significantly changed the field. The interaction of enormous amounts of training data with creative un-supervised or self-supervised learning. The extraordinary successes of modern language models across a wide range of Natural Language Processing (NLP) tasks have been supported by training objectives.

Significant concerns have been expressed in the training community. Pre-trained language models show strong biases that span several dimensions despite their excellent abilities as shown by various research on prejudices impacting marginalised groups, discrepancies in race,sexual orientation and gender.
Due to their potential to have unanticipated and significant effects in practical applications, these biases must be understood. In particular, the prevalence of bias, as demonstrated by instances like the COMPAS AI software employed in US courts, highlights instead of concentrating exclusively on model architecture improvements, address the causes of bias in the training data.To promote fair, responsible language models, it is important to pinpoint the sources of bias within corpora and improve the selection and makeup of training and evaluation datasets that are free of negative prejudices.

In addition, there is a natural bias towards high-resource languages in the field of natural language processing, where they receive disproportionate attention and development ef-

forts. This bias results from the simplicity of data collection, linguistic proficiency, and a continual reinforcement of these languages' development and data gathering.
Despite efforts to close the gap via multilingual language models, there is a widening difference between the amount of text resources accessible for languages with high and low resource levels. Inadvertently minimising the distinctive contributions of under represented languages, multilingual models that rely on skewed language distributions have a tendency to reinforce the perspectives of dominant cultures and languages (Roberto et al.; 2023).

### 2.4 Ethics and Transparency

Large Language Models often operate in a Black Box Model making it challenging to intrepret how a decision is made which can raise ethical concerns. Even some data used to train LLM's have copy wrights leading to infringement and legal complications. Hence ensuring privacy and data protection in the diversified dataset is really important.Identifying the biases and mitigating them is a approach to go forward. All the data must be retrieved with consent and must go through continuous evaluation cycle to identify any ethical and performance issues (Andrés et al.; 2023).

## 3 Methodology

This section focuses on outlining the comprehensive methodology based on the literature review for implementing the proposed solution and conducting the study for this project. The adopted framework experiments with how resources and tools can be efficiently utilized in a resource-constrained environment without compromising the quality of responses. Given the availability of many trained large language models like GPT, T5. In the industry, effectively using them to minimise steps such as purchasing datasets for model training or building highly efficient language models, or hiring human annotators for evaluation can be cost-prohibitive. Therefore, the idea here is to use a pre-trained language model like GPT-2 and fine-tune the existing model to generate domain-specific responses to questions. In this section, we will compare all responses of our fine-tuned model with the baseline GPT model as well as with the T5 language model. The process is divided into sub-sections to provide a detailed context, as also represented in Figure 2:

1. **Research Design** : Transformers have revolutionised NLP and serve as the foundation for many large language models. They excel not only in the efficient processing of sequential data by considering the relationships among all input tokens but also employ a self-attention mechanism that captures context and dependencies. Transformers are widely regarded as highly effective in capturing context and dependencies.
   Pre-trained transformers, when fine-tuned, have the capacity to generate high-quality, coherent, and contextually relevant answers. They can also produce high-quality responses to employee queries, thereby enhancing communication. The design of this study focuses on fine-tuning a pre-trained GPT vanilla model with employee survey data to improve communication within the employee-centric domain, benefiting from its training on extensive data as per (Xin et al.; 2023). The study employs an exploratory approach, combining quantitative and qualitative

measures such as Human Evaluation, Perplexity Calculation, BLEU Score, Average Edit Distance, BERT Score, and METEOR Score. Detailed explanations of these measures are provided in the following paragraphs. Following this, the Google gTTS API is utilised to convert text to speech, considering it as a 'black box',

- **Human Evaluation** : The primary challenge faced by small-scale organisations is the cost associated with hiring human annotators to generate large datasets. This process is not only expensive but also time-consuming, as it requires extensive training. Human annotators are tasked with producing high-quality annotations for data labeling, text annotation, and other tasks. However, due to limited resource availability, the approach here is to conduct human evaluation in terms of Coherence, Relevance, and Overall Quality of the generated responses as suggested by Popović2020. Considering a Rating scale from 1 to 5 for each parameter. With the significance of each column as listed below :
  (i) Input Prompt : The prompt used to generate the response,
  (ii) Generated Response : The response generated by the Fine Tuned GPT2 Model,
  (iii) Coherence : Rating assessment on how the response is logically connected and makes sense,
  (iv) Relevance : Rating assessment to identify how the response is closely related to the input prompt,
  (v) Overall Quality : Considering the above parameters how is the overall response quality is there.
- **Perplexity** : Shannon1948 suggested utilising perplexity, a metric that is well known in literature for the intrinsic evaluation of LMS, to determine how far a real sequence of tokens is from the probability distribution is an effective way for text evaluation,
- **BLEU Score** : As suggested by Kishore et al. 2002 BLEU is not a strong evaluation metric for text generation models, although it can be employed as a scoring function for obtaining quick feedback. In general, it seems that the commonly used automatic metrics heavily emphasize N-gram overlap in a straight forward and simplistic manner. Consequently, if sentences deviate lexically, semantically, or syntactically from the reference phrase, they may not receive recognition, even if they effectively achieve the desired communication objective. Therefore, newer metrics that increasingly employ neural networks to address these issues are widely adopted, in van der Lee Chris et al. (2021) research.
- **Average Edit Distance** : Levenshtein Distance in Levenshtein1965 is also regarded as a significant metric to consider. This edit distance metric measures the dissimilarity between two strings by calculating the minimum number of edits needed to transform one string into the other. Edit distance is commonly employed in fields such as computational biology. A normalised edit distance typically falls within the range of 0 to 1 and elucidates the dissimilarity between the actual and generated text,(Michele and Gianfranco; 2019).
- **BERT Score** : Bidirectional Encoder Representations from Transformers (BERT) in Zhang et al.2020, an NLP method developed at Google, was de-

signed to improve language comprehension tasks. When compared to other NLP tools, BERT has demonstrated superior results across a range of natural language understanding tasks, including paraphrasing and summarisation,in Richard et al. 2021 journal.

- **METEOR Score** : In machine translation research, automatic quality measurements like METEOR from Lavie and Agarwal 2007 are traditionally employed to assess the reliability of Machine Translation (MT) systems, METEOR evaluates translation accuracy by comparing the surface n-grams in the MT output with a human reference. It employs a similar approach to translations that contain one or two grammatical errors, which substantially impact the sentiment of the source text and yield a comparable average score.(Hadeel and Constantin; 2021)

Then techniques like adjusting temperatures, filtering data, top-k sampling, Beam Search, Penalising Repetition were implemented to fine tune the model to achieve best results explained in the next sections,

2. **Data Collection** : A dataset consisting of 20 employee survey responses was collected to support the implementation of the solution. The surveys were designed to capture a range of questions and concerns commonly raised by employees in a business environment. These questions spanned various key areas, including peer learning, team building, aspirations, communication, and more. To safeguard the privacy of the participants and address ethical concerns, the survey results were collected anonymously, with no personal information recorded, not even their IP addresses,

3. **Data Pre-processing** : The survey data underwent several pre-processing steps, which included text normalisation to standardise text entries, data aggregations for organising and categorising employee queries, and data cleaning to remove any inconsistencies or outliers from the data. Some of the steps involved cleaning the data for null and N/A values. For training a language model, the data cannot be directly fed; therefore, tokenisation of sentences, padding for equal length, and then combining them in the data loader were necessary. Once completed, as the initial dataset was retrieved from an Excel file due to its limited size, it was restructured to combine two columns: Questions and Answers. This combined data was then converted into a text file for model training. Similarly, a set of 15 user responses was retained for the training and validation sets. Since the dataset is relatively concise, random up-sampling technique was employed to introduce more diversity into the training set. This process aids in improving the model's performance on unbalanced datasets, which can be enhanced by balancing the class distribution and giving the minority class greater exposure. Additionally, descriptive statistics was used to identify answer lengths, answer counts, the number of unique answers, and other relevant metrics,

4. **NLP Techniques** : Fine-tuning a large language model is a pivotal aspect of this project. Pre-trained language models, one from the Transformers family, are adapted and enhanced using domain-specific data. The objective is to generate answer prompts that are contextually relevant and tailored specifically to employee questions. Achieving this purpose entails fine-tuning and training the models using employee survey data,

5. **Model Training and Evaluation** : The refined model underwent a comprehensive training and evaluation process. Throughout the training, the model's performance was validated using held-out data, involving multiple iterations and parameter optimisation. The evaluation considered criteria such as recall, accuracy, and precision to assess how effectively the models could generate accurate and contextually appropriate responses. Several fine-tuning techniques, as listed below, were selected for model training, including:

   - **Temperature Scaling** : It is a parameter that controls the degree of randomness in the output of a language model. The higher the temperature value, the more diverse the language model's response will be generated. The value ranges from 0 to 1,
   - **Filtering Data** : Filtering data in our case involves removing incomplete data to enhance the quality of the generated data and ensuring consistent token length for text training and generation,
   - **Top-k Sampling** : Koren initially introduced the top-k sampling method as a means to evaluate the performance of top-k recommenders in a seminal study. He conducted a specific comparison between the target movie 'i' and an additional 1000 randomly selected movies. This comparison involved assigning relevance scores, followed by normalizing the ranking values to a range between 0 and 1. Subsequently, cumulative distributions were employed, emphasizing the significance of under-sampling in this context, Dong et al.2020. Hit Ratio and the sampling version are crucial factors in this technique. This approach generates text by emphasizing only the 'k' most probable tokens. The larger the 'k' value, the less biased the result becomes. It resembles a redistribution of probability mass among the 'k' next words,
   - **Beam Search** : It's more of a decoding algorithm aimed at discovering the longest sequence of words for a set of input tokens. It can be viewed as a fragmented version of a graph search, where the potential points are tokens from the input and are systematically retrieved based on an optimisation function. A longer beam length results in more efficient output generation,
   - **Penalising Repetition** : Many researchers have argued that relying solely on fixed corpora cannot fully capture the nuances of a language, and there may be a gap between sampling methods and real human language due to limited corpora in (Zihao et al.; 2021). Hence keeping a repetition threshold in text generation is really important or else the generated text woould not be contextual. All these steps ensure efficient fine tuning of the model is done considering various aspects as well as being ethically responsible.

- **Audio Synthesis** : The study explored audio synthesis as a novel form of communication. It employed a 'black box' approach, treating audio synthesis as a single entity. Although it was not the primary focus, this investigation provided insights into potential future applications of multimedia synthesis in corporate communication,

- **Ethical Consideration** : Throughout the project, ethical considerations held paramount importance. Maintaining anonymity and accuracy were given the highest priority when compiling employee survey results. Stringent data security measures were implemented to safeguard participants privacy, and all ethical standards were meticulously adhered to,

- **Limitations** : The project faces several challenges, including the absence of high-quality domain-specific data for fine-tuning, the complexity of generating and comprehending natural language in a human-like manner, and the inherent complexities of audio synthesis and talking head video technology,

- **Validity and Reliability** : Real-time data collection, process optimization, and model evaluation ensured the project's validity. By maintaining consistency in procedures, transparent reporting, and comprehensive documentation, the study's reliability is demonstrated. To enhance communication in employee-centric corporate environments, the methodology employed a multidimensional approach that integrated NLP techniques and employee survey data. The results, implications, and potential future directions of this comprehensive approach are extensively covered in the following sections.

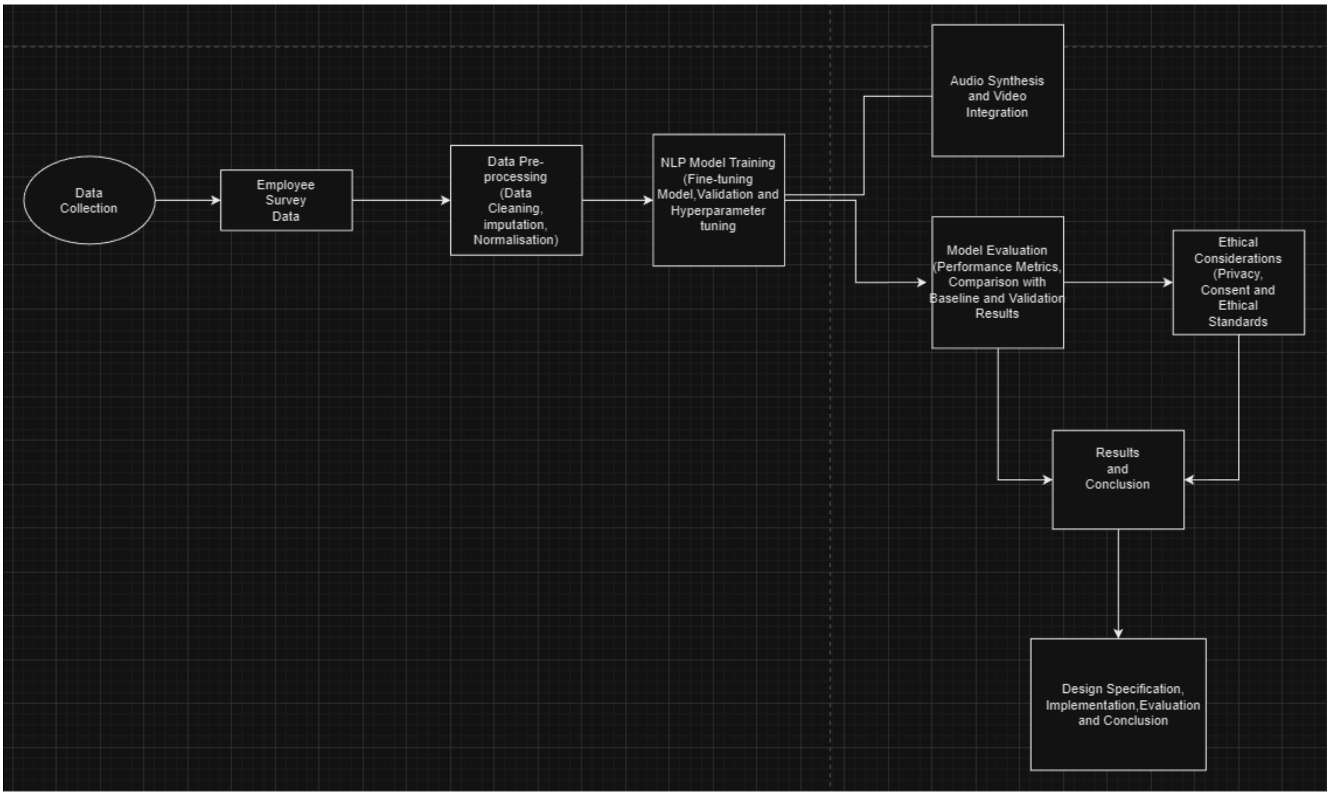



Figure 2: Methodology Flow Diagram

# 4 Design Specification

The detailed specification of the proposed solution, integrating Natural Language Processing, survey data, and multi-media synthesis, which adopts a modular architecture to ensure flexibility and scalability, is explained in this section. Model fine-tuning involves adjusting hyper-parameters to retrieve contextually relevant responses, followed by data processing, including Data Cleaning, Imputation, and Feature Engineering techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) or word embedding. These techniques are used to convert textual survey results into feature vectors. The Fine-Tuned GPT-based transformer model is implemented using Python and the Hugging Face Transformer library, along with the Google Text-to-Speech API for audio generation. The solution also includes robust error-handling components that gracefully manage unforeseen events and maintain user integration.

## 4.1 Model Building/Architecture

The generative pre-trained transformer 2 model is one of the variants of transformers that can be utilised in various tasks like text generation, translation. Fine-tuning of the GPT model simply allows for adjusting model weights to generate responses according to customisation. The architecture of the GPT model is primarily built on top of the transformer model architecture, which involves an encoder and decoder. However, the GPT-2 model falls under auto-regressive language models, focusing on the decoder portion, that is predicting the next word in a sequence of words based on the history of words that came before it, following a feed-forward model approach (Zhan; 2022).

The procedure begins with text provided by the user. To enable the model to process the input text, tokenisation divides the text into smaller units, such as words or subwords. The tokenised text is converted into numerical representations called embeddings at the embedding layer. These embeddings capture the tokens' semantic meaning. A succession of transformer encoder layers are applied to the embeddings. Self-attention mechanisms and feed forward neural networks make up each layer. These layers record the dependencies and contextual ties between the tokens. To provide the model with information about the positions of the tokens in the sequence, positional encoding is added to the embeddings. Each transformer encoder layer's multi-head self-attention mechanisms determine attention weights to weigh the significance of various tokens in relation to each other. To record intricate relationships between tokens, the embeddings undergo a feed forward neural network after self-attention. Then, to stabilize training and promote information flow through the layers, layer normalization and residual connections are used. If necessary, a decoder can be inserted after the encoder for purposes like language generation. Based on the context that the encoder has learned, it produces output tokens. The last layer of the model projects the output of the decoder into a vocabulary space.

Token probabilities are frequently generated using a softmax function. To determine the next token based on expected probabilities during text production, sampling techniques such as top-k sampling or nucleus sampling can be utilised. The generated token(s) are the model's output text. The GPT architecture might be referred to as a "black box" in the sense that it applies intricate calculations and transformations to input text, and it can be difficult to read or explain exactly how it arrives at its output in terms that are understandable to humans. The high-level functionality chart is depicted in Figure 3.

# 5 Implementation

The code implementation begins with the installation of necessary Python libraries, including transformers, GPT-2 Tokeniser, DataLoader, to ensure smooth code execution.

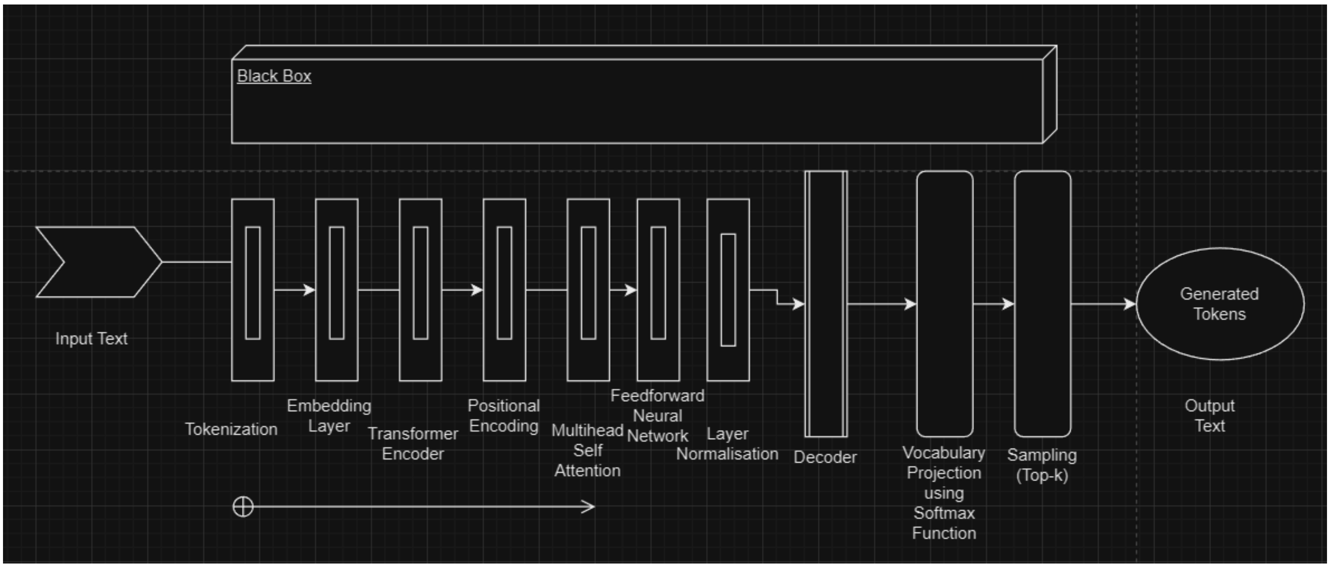



Figure 3: Project Architecture (High Level)

The data is collected through survey responses, as explained in the section above, using SurveyMonkey in a format consisting of questions and free-text answer values. The questions were designed to capture a wide range of ideas from end-users. Proper consent and voluntary participation in the survey were ensured, with no disclosure of personal information, not even the capture of the user's IP address from which the answers were logged. Following this, basic data cleaning is performed, and the data is stored in an Excel sheet format. As part of the pre-processing, tokenisation is applied to the questions and generated answers, followed by padding the sequence length of each sentence and then combining them into a Data Loader. The prepared Data Loader is then fed into the model to generate a comma-separated text file.
The initial GPT-2 model training was run with a block size of 146, and the masked language modeling token was set to False. The processing then moved to the data collator, restricting it to generate responses based on the training dataset. The first fine-tuned model resulted in many repetitions in the generated sentences. Therefore, techniques like setting the repetition threshold to 5 were employed, along with random up-sampling of data to generate diverse responses. As part of multiple model training runs, various parameters were considered for fine-tuning, which are listed below in Table 5

This involved adjusting various parameters, such as modifying the learning rate, fine-tuning the weight decay, changing logging steps to enable more frequent evaluations with the evaluation strategy set as 'STEPS,' utilising a cosine learning rate schedule, and adjusting the gradient accumulation steps for a larger effective batch size, among others. Few parameters were tuned in the model training process. Multiple parameter tweaking was tested, as explained in Section 3, above including top-k sampling, beam search, etc. The number of epochs was set to 5, which is a reasonable value to iterate over the entire training dataset due to its small size. The training batch size was set to 8 to allow more frequent weight updates and require less memory to be processed at once. To ensure stable convergence, the learning rate was kept low at 1e-5, as a higher value was leading to an increase in the loss function value. Lastly, due to the limited volume of data, the data loader worker value was kept at 2.

| Parameters | Value |
|---|---|
| Epochs | 5 |
| Train Batch Size | 8 |
| Save Steps | 10000 |
| Learning Rate | 1e-5 |
| Weight Decay | 0.01 |
| Evaluation Strategy | STEPS |
| Schedular Type | COSINE |
| Gradient Accumulation Steps | 2 |
| Dataloader Workers | 4 |

Table 1: Gpt2 Model Fine-Tuned Parameters

Once the fine-tuned model is created, it is tested on the validation set to determine the total evaluation loss and the number of epochs it executed for, as configured to refine its weights for each run. Subsequently, the model's performance is assessed in comparison to the T5 Language model based on metrics such as Rouge Score, Bert Score, etc., as explained in the sections above. Additionally, Levenshtein Distance is measured as one of the metrics, followed by the creation of visualisations to interpret the results. The tools and configurations used are listed in the configuration manual.

# 6 Evaluation

This part evaluates the performance and effectiveness of the implemented solution, including the components of the tuned model and multimedia synthesis. The evaluation aims to confirm the system's ability to deliver smooth and intuitive user experiences while responding to staff inquiries in a manner that is contextually appropriate and coherent. Later, the gTTS API is used to convert the generated text into an audio file.

## 6.1 Metric Identification

Perplexity, BERT score, and contextual correctness were the main criteria used to evaluate the refined model. While BERT evaluates the quality of generated text in comparison to reference material, perplexity measures the prediction uncertainty of the model. The percentage of responses that are contextually appropriate to the inquiries is measured by contextual correctness. The idea of considering BERT score is that it addresses the limitation of the traditional BLEU score. BERT score functions by generating the F1 score of the generated and reference word comparison, and by taking into account the meaning of words and their relationships within sentences, BERT Score enables a more contextually aware evaluation of text generation, resulting in a more accurate and informative quality assessment (Hadeel and Constantin; 2021). a total of 15 user instances were compared from the Test dataset for these metrics, followed by human evaluation considering various parameters like :

- **Coherence** : Rating 1 to 5 with 5 being logical and well structured flow,
- **Relevance** : Rating 1 to 5 with 5 being highly relevant sentence,

| Questions | Baseline BERT | FineTuned BERT | Baseline METEOR | FineTuned METEOR |
|---|---|---|---|---|
| Question 1 | 0.877262 | 0.887262 | 0.072905 | 0.072905 |
| Question 2 | 0.861597 | 0.871597 | 0.073031 | 0.073031 |
| Question 3 | 0.882704 | 0.882704 | 0.055228 | 0.055228 |
| Question 4 | 0.880993 | 0.880993 | 0.069365 | 0.069365 |
| Question 5 | 0.873708 | 0.893708 | 0.073334 | 0.073334 |
| Question 6 | 0.868394 | 0.888394 | 0.053885 | 0.053885 |
| Question 7 | 0.858131 | 0.898131 | 0.047659 | 0.047659 |
| Question 8 | 0.855239 | 0.865239 | 0.041042 | 0.041042 |
| Question 9 | 0.857462 | 0.877462 | 0.025907 | 0.025907 |

Table 2: Baseline and Fine-Tuned Model Comparison (Question by Question)

| Language Models | Bert Score | Meteor Score |
|---|---|---|
| Fine Tuned GPT 2 | 0.9975467622280121 | 0.9993667128787969 |
| Baseline GPT 2 | 0.9113847613334656 | 0.7311110411437536 |
| T5 Language Model | 0.8481185913085938 | 0.17743709201000377 |

Table 3: Model Final Score Comparison (BERT Score and METEOR Score)

- **Overall Quality** : Rating 1 to 5 with 5 being highly relevant,coherent and of high quality.

The evaluation and comparison are also done with the baseline models, that is Pre-Trained GPT-2 and Text-To-Text Transformer (T5), considering all these metrics.

## 6.2 Evaluation and Comparison

Once the fine-tuned model is trained using the survey response data, it is saved. Following that, text generation for a total of 15 questions is carried out and stored in a data frame from the validation set. Then, the calculation of individual sentence scores for BERT and METEOR is performed, which initially turned out to be the same for the Baseline and Fine-Tuned models, as mentioned in Table 6.2

Hence, based on previous studies, average BERT and Rouge Scores were calculated, including Human Evaluation scores, to conduct a comparative study between the fine-tuned GPT-2 model, GPT-2 model, and T5 Language Model. In this comparison, the fine-tuned GPT-2 model showed quite good results. Here, the fine-tuned GPT-2 model outperformed the T5 model, with the scores as depicted in Table 6.2

The word cloud suggests that the overall survey responses are positive rather than negative (see Figure 6). In addition to it another metric, Average Edit Distance, resulted in a value of 1124.90 Figure 4.

In Figure 5 based on Human Evaluation Considering the length of answer the histogram in Figure 6 suggests that the frequency of questions with a length of 150 or 170 characters is higher compared to others. After building the model and comparing it with the baseline, a few experiments/example scenarios are tested with the fine-tuned GPT-2 model to assess its overall performance.

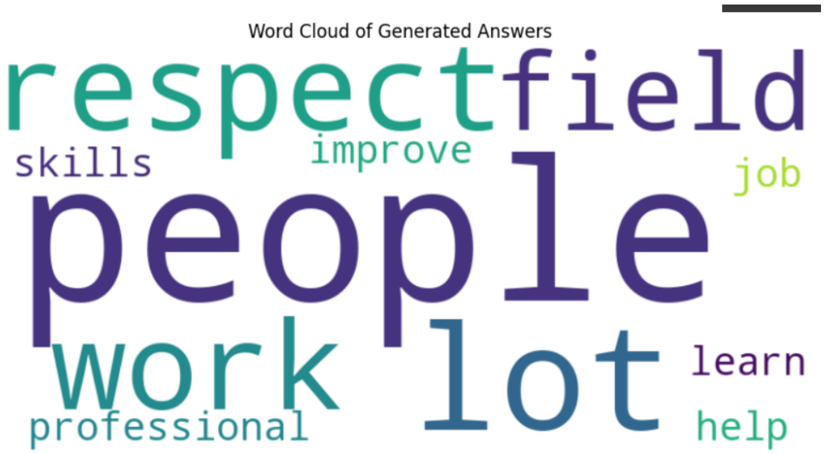



Figure 4: Word Cloud on Survey Responses

| Input Prompt | Generated Response | Coherence | Relevance | Overall Quality |
|---|---|---|---|---|
| What is Peer Learning ? | Peer learning is a process of learning about a group of people. It is a way of learning how to interact with others. Peer learning is about learning how people interact with each other. Peer learners are people who are learning about the world around them. Peer learners are people with whom people share common interests and interests. Peer students are people learning about people who share common experiences and interests, and who share similar interests and experiences. Peer teachers are people in the | 4 | 4 | 5 |
| How team members can collaborate effectively in a corporate environment? | The team members are often the most important part of the team. They are the ones who make decisions, and they are the people who make the decisions. They can be the most effective leaders in a team.<br><br>What are the benefits of working with a team member?<br>.<br>. The team members have the ability to make decisions and make decisions that are beneficial to the team and to the organization.<br>, the team | 2 | 2 | 2 |
| What challenges you face while working in a Team with members having conflicting interest? | I have a lot of different interests. I'm a big fan of the arts and I love the music. I love to work with people who are passionate about their craft. I've been a musician for over 20 years and I've always loved to play music.<br><br>What are your favorite things about working with people with different interests? | 3 | 1 | 2 |

Figure 5: Human Evaluation on Generated Responses

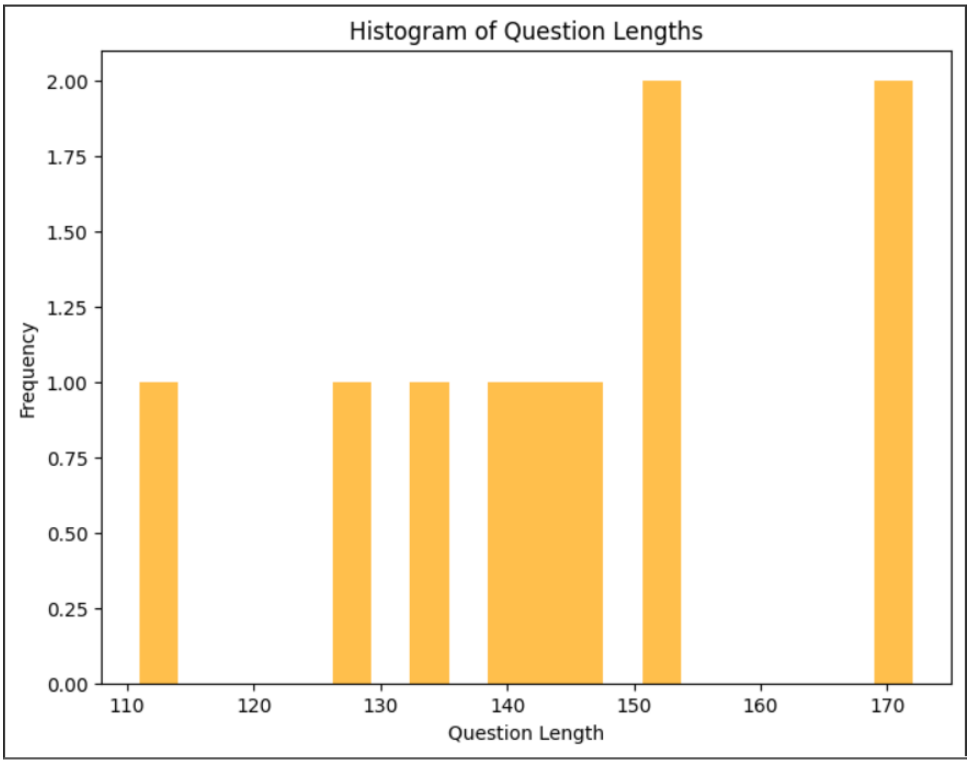
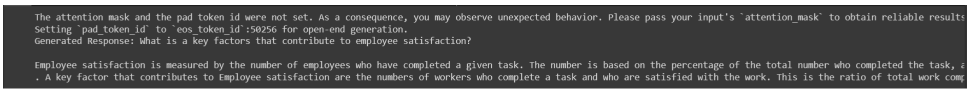



Figure 6: Comparison of Answer Length v/s Frequency

## 6.3 Experiment 1 : Domain Specific Responses

By testing domain-specific queries, we can ensure that the fine-tuned model will be able to provide responses that are accurate and relevant to the specific domain of employee survey data. This demonstrates how the model can indeed be utilized to obtain contextually appropriate answers. Figure 7 showing the snippet from the code and the generated response.



```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask` to obtain reliable results
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Generated Response: What is a key factors that contribute to employee satisfaction?

Employee satisfaction is measured by the number of employees who have completed a given task. The number is based on the percentage of the total number who completed the task,
. A key factor that contributes to Employee satisfaction are the numbers of workers who complete a task and who are satisfied with the work. This is the ratio of total work
```

Figure 7: Domain Specific Responses

## 6.4 Experiment 2 : Contextual Understanding

To ensure that the model can generate coherent and relevant responses in a conversational context, it is evaluated to determine if it can maintain context throughout a conversation. This is crucial for applications that involve back and forth communication. In cases where the model is not generating an appropriate response in the context of the employee's continuation sentence, but instead includes responses in the context of the employer's questions, the contextual performance of the responses can be enhanced by training the model with a larger amount of data. Figure 8 showing the snippet from the code and the generated response.

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask` to obtain reliable results
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Generated Response: Employee: How can I improve my productivity?
Manager: It's important to prioritize tasks and minimize distractions.
Employer: What are your goals for the day? What do you want to accomplish in the future?
```

Figure 8: Contextual Understanding

## 6.5 Experiment 3 : Diversity in Responses

Evaluating the model's ability to generate a variety of relevant responses for the same query. This can be valuable in real-world applications to avoid repetitive or duplicate responses. Figure 9 showing the snippet from the code and the generated response. The

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask` to obtain reliable results
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask` to obtain reliable results
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Diverse Responses:
Response 1: How to be a Team Player?

The best way to be a team player is to be a good teammate. You can't be a good teammate if you don't have a good team
Response 2: How to be a Team Player?

The best way to be a team player is to be a good teammate. You can't be a good teammate if you don't have a good team
```

Figure 9: Diversity in Responses

temperature and top-k parameters were adjusted in an attempt to use the Fine-Tuned GPT-2 model to generate a variety of responses. However, the results did not exhibit the expected level of diversity in the generated responses. The results remained largely consistent across repetitions, despite changing the temperature to encourage randomisation and limiting the tokens taken into consideration using top-k sampling. This unexpected outcome raises questions about the response generation process of the fine-tuned model in general. It is important to note that the challenges in producing genuinely diverse and contextually appropriate responses may be attributed to the inherent properties of the model, in conjunction with the dataset used for fine-tuning. This finding underscores the complexity of response generation tasks, emphasising the need for further research and testing to achieve the desired output.

## 6.6 Experiment 4 : Transfer Learning

A scenario simulating the model's need to adapt to a new context or domain is created by fine-tuning the model using a different dataset. Therefore, evaluating a model's ability to generalise and adapt by testing its performance on such data is one of the crucial steps. Figure 10 showing the snippet from the code and the generated response.

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask` to obtain reliable results
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask` to obtain reliable results
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Generated Response: Employee feedback: More one to one meetings should be organised.

The first meeting should include:
...
,
:.,
 (or)
(or).. (Or).(Or).
.. (Not) (not). (No) or..(Not).(Not.)
"

Generated Response: Employee feedback: New events needs to be added in the calendar.

New events needed to add in. New event management: Event management should be integrated into the event system. This should allow for better integration of events into events ma
```

Figure 10: Diversity in Responses

The use of transfer learning in this study has yielded promising results. By fine-tuning a large pre-trained language model, particularly the GPT-2 model, with a domain-specific dataset comprising employee survey responses, the model demonstrated a remarkable ability to generate contextually appropriate and coherent responses to various questions. The refined model exhibited an improved understanding of domain-specific terminology and nuances, enabling it to produce responses closely aligned with the context of employee survey questions. This demonstrates that transfer learning offers an effective method to leverage the extensive linguistic knowledge of existing language models and adapt them to domain-specific tasks.

# 7 Conclusion and Future Work

To conclude, this work set out on a mission to harness the potential of transfer learning in the field of natural language generation, focusing especially on response generation for domain-specific queries connected to peer learning in the Human Resources context. We have discovered a number of observations and positive results through the fine-tuning of the GPT-2 language model using a dataset made up of employee survey responses.
The ability of the GPT-2 model to produce contextually coherent responses in line with the subtleties of employee survey queries is considerably enhanced by the application of transfer learning, as demonstrated by our studies in the first place. The improved model showed a significant improvement in understanding domain-specific vocabulary and successfully adapting its language to address HR-related queries.

The results highlight the potential of transfer learning as a game-changing tool for adapting existing language models to task-specific domains, making significant advances in the area of natural language processing. Considering the results from the study, and acknowledging the limitations is really important, as the solution provided by the refined model may not fully account for the complexities, and further work needs to be done to improve the model's capabilities in the comprehension of complicated situations. Additionally, before actual deployment, ethical considerations and potential biases in the created information should be carefully examined and addressed, as with any machine learning approach. In summary, this research expanded our understanding of how transfer learning might enhance response production in specialised fields and opens new avenues for more effective and contextually relevant answer generation in a resource-constrained situation.

## 7.1 Future Work

Building on the successes of this work, a number of exciting lines of enquiry for future study emerge that can deepen our understanding and push the boundaries of response creation within and outside of Human Resource Domain.

- Diversity in Responses: To address the issue of response diversity, future studies should explore sophisticated sampling techniques such as nucleus sampling to encourage a more varied and contextually relevant range of responses, accommodating a wider range of HR circumstances,

- Hybrid Approaches: To strike a balance between regulated response creation and the contextual richness offered by refined models, future research may explore the combination of rule-based systems with various large language models and diversified datasets,

- Collaboration between humans and Artificial Intelligence (AI) holds enormous potential, and research can be conducted to determine if the improved model can assist HR professionals by engaging text annotators,

- Transfer to Other Domains: Investigating how the improved model can be applied to other specialised domains, such as creating a cross-lingual platform for text generation and using automatic *Machine Translation* (MT) to translate prompts into other native languages, which can be of assistance in various domains like aiding people during crises such as earthquakes and cyclones, presents an intriguing opportunity to enhance the model's utility.

We have gained insight into the dynamic world of AI-powered response generation through our exploration of the areas of fine-tuning and transfer learning. These findings lay the foundation for future research and significant progress in the rapidly evolving fields of natural language processing and AI-driven communication.

# A Appendix

The collaboration with the company Hyworx, which played a crucial role in providing access to real-time staff survey data, was the driving force behind this study effort. The motivation behind this study stemmed from Hyworx's commitment to promoting a healthy work environment and their interest in using technology to enhance employee engagement. This partnership enabled the development of cutting-edge applications of AI-powered language models to address issues in employee communication and interaction. This dataset served as the foundation for the research, enabling the investigation.In addition to it Figure 11 lists the brief overview on the survey dataset.

| **Questions** | **User1** | **User2** | **User3** | **User4** | **User5** |
|---|---|---|---|---|---|
| **How do you approach peer learning to make the most out of the experience?** | I actively participate in group discussions, ask clarifying questions, and take notes to retain important information. | I like to share my knowledge and expertise with others while being open to learning from my peers' unique perspectives. | I believe in setting clear learning goals with my peers and collaborating on projects that challenge and inspire us. | I engage in regular peer reviews and seek feedback to improve my understanding and learn from different viewpoints. | I find it helpful to form study groups and schedule regular meetings to stay accountable and on track with my peers. |
| **What challenges have you encountered while engaging in peer learning , and how did you overcome them?** | One challenge I faced was coordinating schedules with different group members. We used online scheduling tools to find common meeting times and set a fixed weekly session. | Sometimes, there were varying levels of knowledge among peers, which made discussions challenging. To overcome this, we organized peer mentoring sessions, where experienced peers guided | I encountered shyness and hesitancy to ask questions in the group. We implemented ice-breakers and team-building activities to create a supportive and inclusive atmosphere. | Maintaining equal participation among all group members was difficult. We introduced a rotation system for presenting topics, ensuring everyone had an opportunity to contribute. | We faced occasional conflicts over differing opinions. To address this, we practiced active listening and encouraged open discussions to find common ground. |
| **Share a memorable experience where peer learning made a significant impact on your learning journey?** | During a difficult math problem, my peer offered a unique approach that helped me understand the concept better, leading to my improved grades | Working on a group project, my peers challenged me to think critically, pushing me beyond my comfort zone and enhancing my problem-solving | Through peer learning, I gained insights into real-life applications of theoretical concepts, making my learning more practical and | A peer shared their personal experiences in the field, inspiring me to pursue a related career path and shaping my long-term goals | In a study group, my peers and I prepared mini-presentations on different topics, and it broadened our knowledge by learning from |

| **Questions** | **User6** | **User7** | **User8** | **User9** | **User10** |
|---|---|---|---|---|---|
| **How do you approach peer learning to make the most out of the experience?** | I take the initiative to organize mini-presentations on topics I excel in, promoting a dynamic learning environment. | I actively seek out peers who have expertise in areas I want to learn more about and engage in one-on-one discussions. | I embrace diversity in my peer group and encourage open discussions, fostering a rich exchange of ideas and knowledge. | I use online collaborative tools and forums to share resources and engage in discussions beyond regular study sessions. | I appreciate the value of active listening, as it allows me to absorb and respect my peer contributions. |
| **What challenges have you encountered while engaging in peer learning , and how did you overcome them?** | Time management was a challenge, especially during busy periods. To overcome it, we set specific learning goals for each session and adhered to a structured agenda. | Some members struggled with the online learning platform we used. We offered technical support and provided tutorials to familiarize them with the tools. | Language barriers occasionally hindered effective communication. We encouraged the use of simpler language and provided translations when necessary. | Maintaining motivation during long study sessions was tough. We introduced short breaks and incorporated fun activities to keep the energy high. | We faced challenges in coordinating group assignments, leading to delay To improve efficiency, we established a shared virtual workspace for collaboration. |
| **Share a memorable experience where peer learning made a significant impact on your learning journey?** | During a debate, a peer presented a perspective I had never considered, expanding my understanding and enriching my learning journey | My peer introduced me to online resources and study techniques that significantly improved my study habits and academic | Collaborating with peers from diverse backgrounds exposed me to different cultures and ways of thinking, fostering a global | In a group discussion, a peer asked insightful questions that prompted deeper exploration of the subject matter, leading to a more | A study partner provided emotional support during stressful times, making the learning journey less daunting ar more enjoyable |

Figure 11: Survey Dataset

# References

Andrés, P.-M., Carmen, G.-M., Laura, D.-F. and del Carmen, L.-P. M. (2023). Ethical challenges in the development of virtual assistants powered by large language models, *Electronics* **12**(14): 3170.

Dong, L., Ruoming, J., Jing, G. and Zhi, L. (2020). On sampling top-k recommendation evaluation, *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2114–2124.

G, M. J., J, U. R., CN, M. P., Karen, O., Ruowang, L., Pei-Chen, P., J, B. T., Nicholas, T., Jae, W. K., Graciela, G.-H. et al. (2023). Chatgpt and large language models in academia: opportunities and challenges, *BioData Mining* **16**(1): 20.

Hadeel, S. and Constantin, O. (2021). Bleu, meteor, bertscore: evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text, *arXiv preprint arXiv:2109.14250* .

Jezia, Z., Mohamed, A. J. and Imad, S. (2022). Domain-specific text generation for arabic text summarization, *2022 International Conference on Computer and Applications (ICCA)*, IEEE, pp. 1–4.

Kishore, P., Salim, R., Todd, W. and Wei-Jing, Z. (2002). Bleu: a method for automatic evaluation of machine translation, *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318.

Lavie, A. and Agarwal, A. (2007). Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments, *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 228–231.

Levenshtein, V. (1965). Contribution to levenshtein distance/introduction of levenshtein distance concept, Personal communication / Pioneering contribution.

Michele, S. and Gianfranco, B. (2019). Bounds and estimates on the average edit distance, *International Symposium on String Processing and Information Retrieval*, Springer, pp. 91–106.

Moslem, Y., Haque, R. and Way, A. (2023). Adaptive machine translation with large language models, *arXiv preprint arXiv:2301.13294* .

Popović, M. (2020). Informative manual evaluation of machine translation output.

Richard, C. C., Charuta, P. and Steven, S. (2021). Natural language processing versus rule-based text analysis: Comparing bert score and readability indices to predict crowdfunding outcomes, *Journal of Business Venturing Insights* **16**: e00276.

Roberto, N., Simone, C. and Björn, R. (2023). Biases in large language models: Origins, inventory and discussion, *ACM Journal of Data and Information Quality* .

Shannon, C. E. (1948). A mathematical theory of communication, *The Bell System Technical Journal* **27**(3): 379–423.

Tibor, K. Z., Tibor, C., G, P. B. S., Szilárd, B., Ildiko, N.-V. and Andrea, B. (2023). Measuring organizational happiness in higher education-a data-driven approach, *Cogent Education* **10**(1): 2190382.

van der Lee Chris, Albert, G., van Miltenburg Emiel and Emiel, K. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines, *Computer Speech & Language* **67**: 101151.

Xin, Z. W., Kun, Z., Junyi, L., Tianyi, T., Xiaolei, W., Yupeng, H., Yingqian, M., Beichen, Z., Junjie, Z., Zican, D. et al. (2023). A survey of large language models, *arXiv preprint arXiv:2303.18223* .

Yu, Y., Wenzhuo, D., Di, Y. and Dechun, Y. (2023). Cipta: Contrastive-based iterative prompt-tuning using text annotation from large language models, *2023 4th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)*, IEEE, pp. 174–178.

Zhan, J. (2022). A benchcouncil view on benchmarking emerging and future computing.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1499–1515.

Zihao, F., Wai, L., Man-Cho, S. A. and Bei, S. (2021). A theoretical analysis of the repetition problem in text generation, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 12848–12856.