

AIML B20 - PGCP

# AI-based Generative QA System

**Group -3**

## Group Member Name 1 Group Member Name 2 Group Member Name 3 Group Member Name 4

**Project Mentor:**  Sagar

**Project Supervisor:**  Manish Shrivastava

## June 11, 2023

1. **Introduction 3**
2. **Email Subject Line Generation (Task 1) 3** 
   1. Literature Review 3
   2. Proposed Approach 3
      1. Understanding the Dataset 3
      2. Model 4

GPT-2 4

* + 1. Model Training & Deployment 4
  1. Applicability in the real world 5

1. **Question Answering on AIML Queries (Task 2) 5** 
   1. Literature Review 6
   2. Dataset 6
   3. Proposed Approach 6
   4. Applicability in the real world 6

E-Commerce: 6

E-tutor for students: 7

Government Policies: 7

1. **Challenges 7**
2. **Project Timeline 8**
3. References 8

### 1. Introduction

This document proposes the plan for the two given tasks of the project. The objective is to learn and make use of available Generative AI models.

### 2. Email Subject Line Generation (Task 1)

The uniqueness of the task lies in having to generate an extremely short, concise summary in the form of the email subject. This involves identifying the most salient sentences from the email body, and abstracting the message contained in those sentences into only a few words.

#### 2.1. Literature Review

While a lot of work has been carried out for news headline generation, news and email summarization, email-subject line generation was first introduced by  [Zhang and Tetreault, 2019 .](https://arxiv.org/pdf/1906.03497.pdf) The paper describes their contribution, which is threefold:

1. introduced the task of email subject line generation (SLG) and built a benchmark dataset  [AESLC](https://github.com/ryanzhumich/AESLC) .
2. investigated possible automatic metrics for SLG and studied their correlations with human judgments. They also introduce a new email subject quality estimation metric (ESQE).
3. proposed a novel model to generate email subjects.

The work of  [Zhang and Tetreault, 2019](https://arxiv.org/pdf/1906.03497.pdf) is inspired by  [Chen and Bansal, 2018.](https://aclanthology.org/P18-1063.pdf) However, the major difference between the two is:  [Chen and Bansal, 2018](https://aclanthology.org/P18-1063.pdf) assumed that there is a one-to-one relationship between the summary sentence and the document sentence. In contrast,  [Zhang and Tetreault, 2019](https://arxiv.org/pdf/1906.03497.pdf)  extracted multiple sentences and rewrote these sentences together into a single subject line.

#### 2.2. Proposed Approach

##### 2.2.1. Understanding the Dataset

The  [AESLC](https://github.com/ryanzhumich/AESLC)  dataset was prepared by  [Zhang and Tetreault, 2019](https://arxiv.org/pdf/1906.03497.pdf)  where they use the Enron dataset, which is a collection of email messages of employees in the Enron Corporation. The Enron data set contains both business and personal type emails. In the  [AESLC](https://github.com/ryanzhumich/AESLC) dataset, the average number of words in a document is 75 and the average number of summary words is 4. The total number of documents is 14,436, out of which 1 ,960 can be used for validation, 1 ,906 for testing, and the rest for training.

##### 2.2.2. Model

Various trained transformer models are openly available on

[HuggingFace .](https://huggingface.co/) We propose to use the same models and fine tune them on our data-set. Some of the models are:

*GPT-2*

[GPT-2](https://openai.com/research/better-language-models) is a transformer model pre-trained on a very large corpus of English data in a self-supervised fashion. It has been pre-trained on the raw texts only, with no human labeling, and with an automatic process to generate inputs and labels from those texts. GPT-2 was first proposed by  [Radford et. al. 2019](https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf) , later on  [Budzianowski and Vulic, 2019](https://aclanthology.org/D19-5602.pdf)  have also proposed a method to tune and use GPT-2 for a specific domain, On  [HuggingFace ,](https://huggingface.co/) three other variants of GPT-2 are available (as listed in Table 1), depending on availability of hardware resources, one of these variants will be used.

|  |  |
| --- | --- |
| **GPT-2 variant** | **Number of Parameters** |
| GPT-2 Large | 774 Million |
| GPT-2 Medium | 355 Million |
| GPT-2 XL | 1.5 Billion |

**Table 1** : Different variants of GPT-2 on  [HuggingFace](https://huggingface.co/)

##### 2.2.3. Model Training & Deployment

Model training would be carried out in three typical stages (as shown in Figure 1).

Data preprocessing involves tokenizing the data, which is the process of converting a sequence of characters into tokens, i.e. separating a sentence into words.

Training model involves importing the pretrained GPT-2 model, as well as the tokenizer. Since, GPT-2 has a lot of parameters. It may be possible that we face  *OUT OF MEMORY*  errors. Hence, as an alternative, we plan to accumulate the gradients. The idea is that before calling for optimization to perform a step of gradient descent, we will attempt to sum the gradients of several operations. Then, divide that total by the number of accumulated steps, in order to get an average loss over the training sample. This method would reduce the number of calculations.



**Figure 1** : Training the GPT-2 model

Performance evaluation would be carried out with METEOR, ROUGE or BLEU matrices, depending on suitability.

The model will be deployed in a Web-based interface using FASTAPI.

Some of the tools required during the project would be as follows -

**Tools** : Hugging Face, PyTorch, Tensorflow, Keras, WandB, NLTK

**Deployments:**  FastAPI, Cloud Application Platform | Heroku, Streamlit,

Cloud Computing, Hosting Services, and APIs | Google Cloud

#### 2.3. Applicability in the real world

The application of email-subject line generation is immense, especially for the commercial email service providers. This technology would help users to save time. Generated subject lines would help users in categorizing and searching emails more efficiently.

Another application of this technique would be assisting visually impaired and physically disabled users, where they would just type or dictate the email without worrying about its subject.

### 3. Question Answering on AIML Queries (Task 2)

The problem statement is to build a Question & Answering system for queries related to AIML, this system will be designed to process queries in natural language and provide correct and relevant solutions.

This task involves modeling a domain-specific GPT-variant model that can answer the questions specific to the AIML course. Pre-trained models can produce relevant textual output for general, open-domain textual prompts, however, these models lack the capability of producing finer outputs when it comes to domain-specific tasks. Therefore, similar to Task 1 (as proposed in 2.2 Methodology ), a pre-trained model will be finetuned, to tailor the model’s expertise.

#### 3.1. Literature Review

The GPT model was proposed by  [Radford et. al,](https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf)  It is a causal (unidirectional) transformer pre-trained using language modeling on a large corpus of long range dependencies, the Toronto Book Corpus. Later, various improvements were made, and subsequently,  [GPT-2 ,](https://openai.com/research/gpt-2-1-5b-release)  [GPT-3](https://openai.com/blog/gpt-3-apps/) and  [GPT-4](https://openai.com/research/gpt-4) have been released.

#### 3.2. Dataset

A custom dataset consisting of various queries and their responses will be prepared. This dataset would also contain multiple queries for a response and various responses for a query. A sample of query and response is given in Table

2.

|  |  |  |
| --- | --- | --- |
| **Sample Query 1** | **Sample Query 2** | **Sample Response** |
| What is  Stocashtic Gradient  Descent? | What happens when I use SGD? | Stochastic Gradient Descent (SGD) is an iterative method for optimizing an objective function with suitable smoothness properties. It can be regarded as a stochastic approximation of gradient descent optimization, since it replaces the actual gradient by an estimate thereof. This reduces the computational burden, especially in high-dimensional optimization problems. |
| Tell me about Random  Forest | What is  Random  Forest? | Random forest is a supervised machine learning algorithm that can be used for both classification and regression problems 1. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. |

**Table 2** : Sample Queries and Response

#### 3.3. Proposed Approach

The proposed approach to the problem is the same as proposed for Task-1 in Section 2.2 .

**3.4. Applicability in the real world**  *E-Commerce:*

This system would be of great use for online shopping, as the customer can get their queries solved regarding a product and can review before ordering. This would result in less number of returned products and customer satisfaction.

*E-tutor for students:*

Many times, students have common queries in a subject, using Generative AI and training it on a custom dataset (similar to Task 2) would help students get their doubts solved at any time.

*Government Policies:*

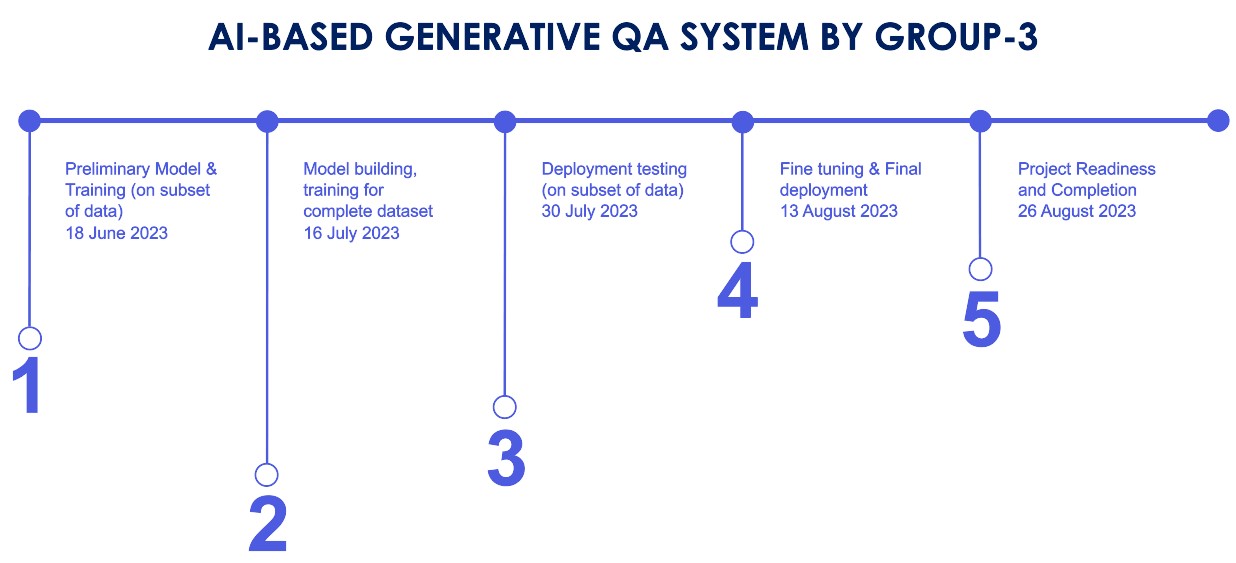
Although governments offer many policies for common citizens, the utilization of these policies is often unused or misused. Having a virtual assistant which can solve all queries related to government policies would be of great help to common citizens in creating awareness.

### 4. Challenges

Some of the challenges we foresee during the development for Task 1 and Task 2 are as follows -

* Monitoring and continuous improvement to track the performance and usage of the QAS.
* Continuously gather user feedback and literate on the system to address any improvement.
* Lack of Diversity in Training Data
  1. The QA system may struggle to understand and respond to certain types of user input. For example, if the training data is biased towards a particular demographic or language style, the chatbot may struggle to understand and respond to users who do not fit that demographic or language style. This could lead to frustration for users and a poor user experience overall.
* Overfitting in GPT Prompt Training
* Difficulty in Capturing Contextual Nuances
* Bias and misinformation
  1. The dataset used for training the GPT model shall be unbiased and shall have accurate information otherwise generating the QA system would be producing biased or inaccurate responses.

### 5. Project Timeline



### 6. References

1. [https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language](https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf)

[-unsupervised/language\_understanding\_paper.pdf](https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf)

1. <https://huggingface.co/docs/transformers/model_doc/openai-gpt>
2. models/language\_models\_are\_unsupervised\_multitask\_learners.pdf
3. Zhang, R., & Tetreault, J. (2019, June 8).  *This email could save your life: Introducing the task of email subject line generation*. arXiv.org. <https://arxiv.org/abs/1906.03497>
4. Ryanzhumich. (n.d.). GitHub - ryanzhumich/AESLC: Annotated Enron Subject Line Corpus (AESLC). GitHub. <https://github.com/ryanzhumich/AESLC>
5. <https://www.cs.cmu.edu/~enron/>
6. https://aclanthology.org/D19-5602.pdf