## **INDIGO CASE STUDY**

## **Introduction**

**Question Answering (QA) systems** have gained significant traction in recent years, with applications ranging from customer support to information retrieval. This project aims to develop a robust QA model capable of generating accurate and informative responses to a wide range of user queries. By leveraging the extensive Quora Question Answer dataset, I seek to create a system that emulates human-like interaction and provides valuable insights.

This report outlines the development and evaluation of a question-answering model, exploring data preprocessing techniques, model architecture, and performance metrics to achieve optimal results.

## **Literature Survey**

**Question Answering (QA)** has witnessed significant advancements with the advent of deep learning. Early approaches relied on rule-based systems and information retrieval techniques, often falling short in handling complex queries. The introduction of Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) marked a step forward, but their limitations in capturing long-range dependencies persisted.

Attention mechanisms, as proposed by Bahdanau et al. (2014), addressed these limitations by allowing models to focus on relevant parts of the input. This led to improved performance in various NLP tasks, including QA.

Transformer-based models, such as **BERT** (Devlin et al., 2018), **T5** (Raffel et al., 2019), and **GPT** (Radford et al., 2018), have revolutionized the field. BERT, a bidirectional encoder, excels in capturing contextual information, making it suitable for QA tasks. T5, a text-to-text transfer transformer, offers flexibility in handling various NLP tasks, including QA. GPT, primarily a generative model, has shown promising results in open-ended QA scenarios.

While these models have achieved state-of-the-art results on various benchmarks, challenges such as handling complex queries, factual inconsistencies, and generating diverse and informative answers remain. This project aims to contribute to the ongoing research by comparing the performance of BERT, T5, and GPT on the Quora Question Answer dataset.

## **Methodology**

### **Dataset Description**

The Quora Question Answer Dataset is designed to develop and evaluate question-answering models. The dataset consists of pairs of questions and answers, sourced from Quora, and is suitable for natural language processing tasks such as training models to generate accurate responses to user queries.

question: A string representing the question asked.

answer: A string representing the answer to the question.

Dataset Size:

Total entries: 56,402

Training set size: 45,121 (80% of the total data)

Validation set size: 11,281 (20% of the total data)

### **Data Preprocessing**

To prepare the data for model training, several preprocessing steps were undertaken:

* **Data Cleaning:**
  + Removed irrelevant HTML tags, emojis, and \n.
  + Addressed spelling errors and inconsistencies.
* **Exploratory Data Analysis (EDA):**
  + Conducted univariate analysis on text length, word frequency and stop words.

1. Distribution of Text Lengths
2. Question

**(i) Majority of questions are relatively short:** This indicates a preference for concise and direct queries.

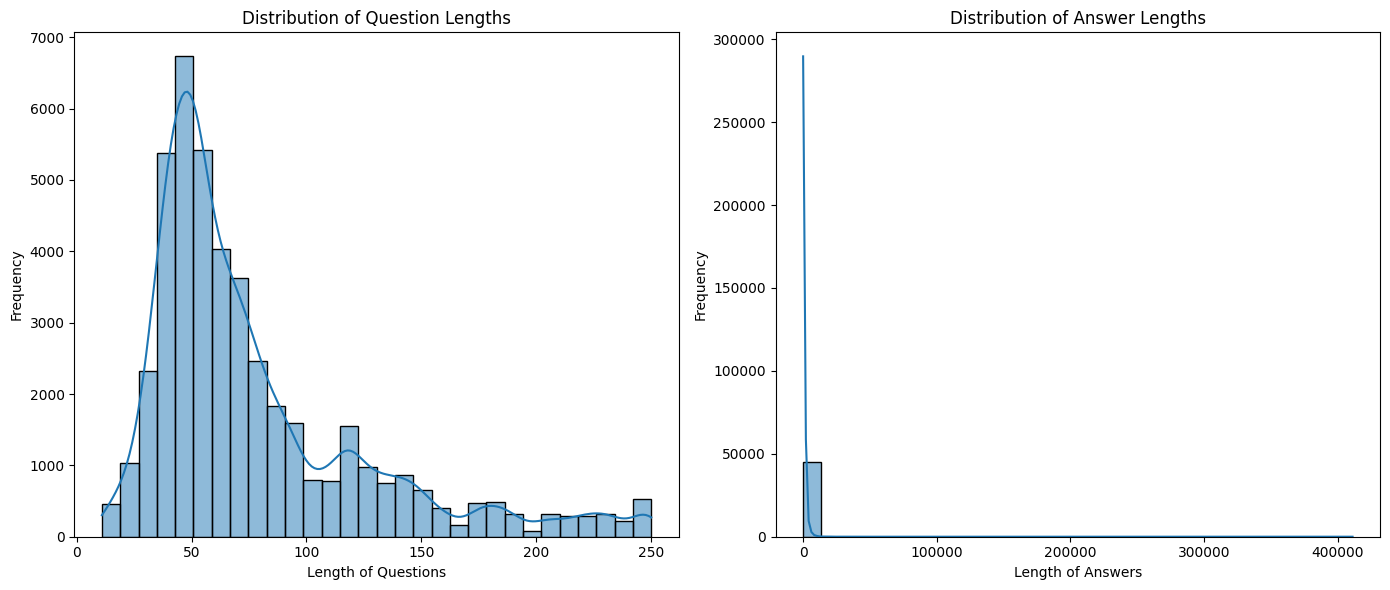
**(ii) Distribution is right-skewed:** There are a few instances of longer questions, but they are less frequent.

1. Answer

**(i) Answers tend to be longer than questions:** This suggests that users often provide detailed explanations or responses.

**(ii) Distribution is also right-skewed:** Similar to questions, there are fewer extremely long answers.

**(iii) Longer tail compared to questions:** The distribution shows a larger proportion of longer answers than longer questions, indicating a wider range of answer lengths.



2) Column Word Analysis

1. Question

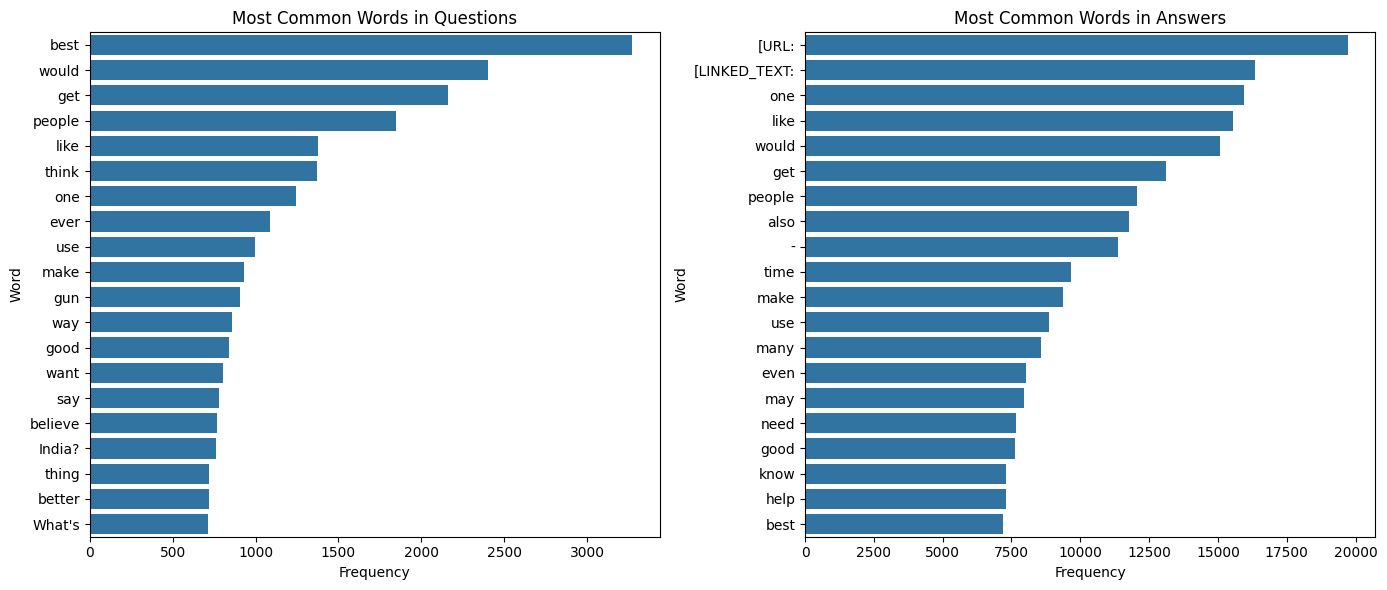
**(i) Focus on information seeking:** Words like "what", "who", "when", "where", "why", and "how" are frequent, indicating a strong focus on gathering information.

**(ii) Open ended questions:** The presence of words like "best", "like", "think", and "one" suggests open ended questions that seek opinions or explanations.

1. Answer

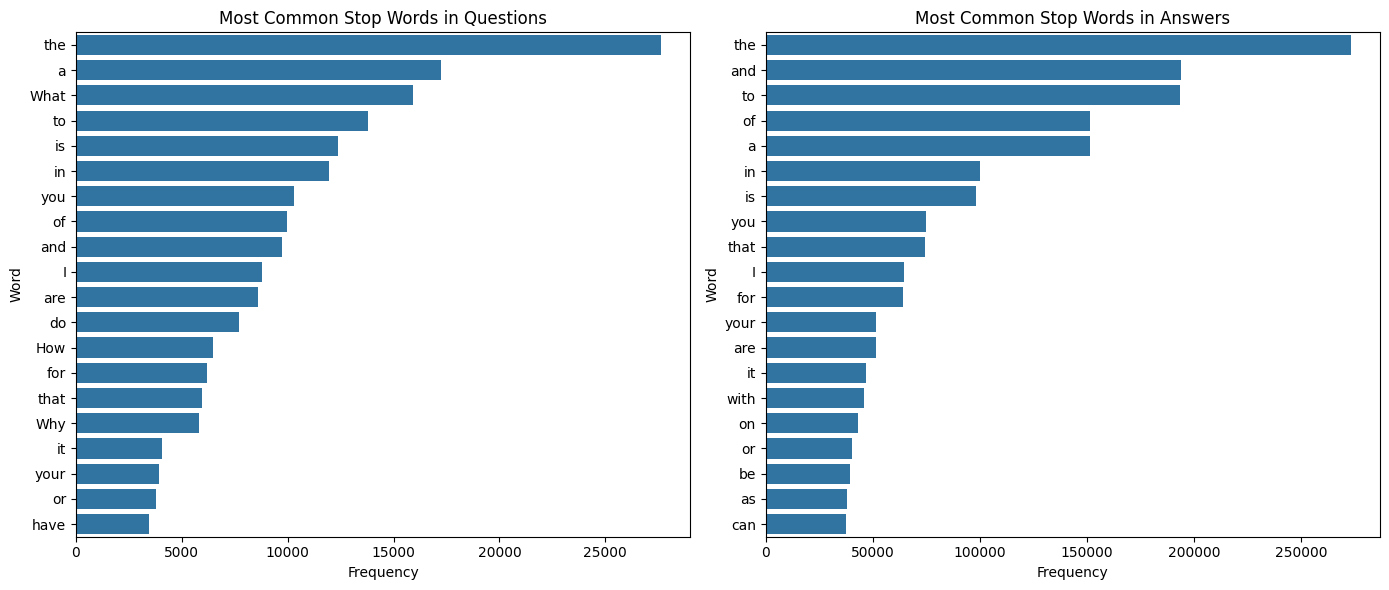
**(i) Informative responses:** Words like "get", "use", "make", "good", "way", "time", "know", "help", and "need" indicate informative answers that provide solutions or explanations.

**(ii) Varied vocabulary:** Compared to questions, answers seem to have a more diverse range of words, possibly because they elaborate on concepts and provide specific details.





2) Stop Word Analysis



### **Model Selection**

I have selected three state-of-the-art language models for comparison: BERT, T5, and GPT. These models have demonstrated exceptional performance in various NLP tasks, including question answering.

### **Model Training**

1. BERT Model

**Model Architecture:** The pre-trained bert-base-uncased model was employed. This model features 12 encoder layers, 768 hidden dimensions, and a dedicated output layer for question answering.

**Training Data:** The preprocessed Quora dataset was divided into training and validation sets.

**Training Process:** The model underwent training for 2 epochs with a batch size of 16. The AdamW optimizer and a learning rate scheduler were utilized for optimization. During training, batches were processed, and the model's loss was computed based on its output for question-answer pairs. The optimizer subsequently updated the model parameters to minimize the loss.

1. GPT Model

**Model Architecture:** The pre-trained gpt2 model was employed. This model is adept at predicting the next word in a sequence.

**Training Data:** The preprocessed Quora dataset was divided into training and validation sets.

**Training Process:** The model was fine-tuned on the training data for 2 epochs with a batch size of 16. During training, batches of text sequences were processed, and the model's loss was calculated based on its ability to predict the next word in the sequence. The training objective is to steer the model towards generating answers that follow question prompts within the training data.

1. T5 Model

**Model Architecture:** The pre-trained t5-small model was employed. This model features an encoder-decoder structure with 6 encoder and decoder layers.

**Training Data:** The preprocessed Quora dataset was divided into training and validation sets.

**Training Process:** The model underwent fine-tuning on the training data for 3 epochs with a batch size of 2. During training, batches of text sequences were processed, and the model's loss was calculated based on its ability to predict the appropriate response to the provided question (essentially, a translation task).

### **Model Evaluation**

The performance of the models will be evaluated using the following metrics:

* ROUGE: To assess the overlap between generated and reference answers.
* BLEU: To evaluate the fluency and correctness of generated answers.
* F1-score: To measure the overall performance of the models.

| **Model** | **BLEU Score** | **ROUGE-L Score** | **F1 Score** |
| --- | --- | --- | --- |
| **BERT** | 0.006863021393776 | 0.010821782976539 | 0.01093706833426 |
| **GPT** | 0.000198811738394 | 0.04767386358125 | 0.0416320263505 |
| **T5** | 0.000153672700927 | 0.04644996788776 | 0.0467295369556 |

## **Conclusion**

This report compared the performance of BERT, GPT, and T5 models in generating answers to questions using the Quora Question Answer dataset.While both models were able to generate some relevant responses, the evaluation metrics (BLEU, ROUGE-L, and F1) indicate that the GPT model demonstrated superior performance.

## **Novel Improvements**

### **1. Hybrid Model Architecture:**

* **Combine the strengths of BERT and GPT :** Leverage BERT's strong contextual understanding capabilities with GPT's generative fluency.
* **Create a hierarchical model:** Employ BERT for initial question comprehension and information retrieval, followed by GPT for generating detailed and coherent answers.

### **2. Evaluation Metric Enhancement**

* **Semantic Similarity Metrics:** Develop more sophisticated metrics that assess the semantic similarity between generated and reference answers, going beyond syntactic overlap.