NEWS ARTICLE BASED QUESTION ANSWERING SYSTEM

Presented By: Sushovan Saha
214161010,
M.Tech-Data Science

SUPERVISORS: Dr. Ashish Anand
& Dr. Prithwijit Guha

OUTLINE

1. Introduction
2. Motivation
3. Literature Review
4. Contributions
5. Experiments and Results
6. Future Work and Conclusions

INTRODUCTION

- The task of **News Question Answering** (**NQA**) is to generate **semantically** and **syntactically** correct natural language responses to questions that are related to recent or ongoing news events.

- Example :

  Q : Which company clocked 2 billion in GMV in calendar 2020?

  A : Myntra

MOTIVATION

- If an Investigation is going on some historical events.
- Manually searching for answer related to past events from news archive is a cumbersome process.
- Questions that cannot be answered using synchronic data source like Wikipedia

LITERATURE REVIEW

- A large amount of unstructured data is required for tasks like Open Domain Question Answering (ODQA).

- ❑ **SQuAD Dataset** : The SQuAD database is the most well-known data collection for the Question Answering activity. Wikipedia articles served as the data set's primary source.

- ❑ **NewsQA Dataset :** Our work's most relevant data collection is NewsQA. It has 12,744 CNN news stories and 119,633 natural language questions assembled by crowd workers.

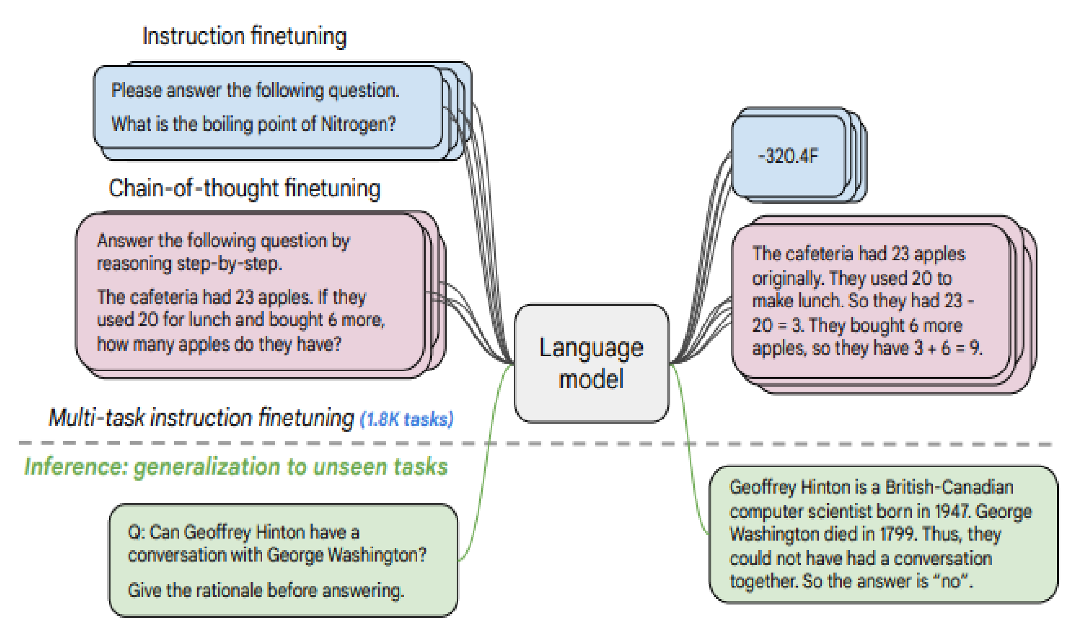
LITERATURE REVIEW

- **ArchivalQA** : NewsQA is adequate, but their curation is done by crowdsourcing, which is a laborious procedure in and of itself. A framework for building data sets is offered in ArchivalQA and may be used to create a data set on a temporal collection of articles.

- **BERTserini** : It is a retriever reader ODQA model. Anserini is a single-stage retriever that can identify segments of text from the material. The BERT reader selects the best text span and helps to get the answer.

LITERATURE REVIEW

- **Text-to-Text Transfer Transformer (T5) :** It is created by Google researchers, can combine several activities under a single framework. T5 model performs exceptionally well in tasks including text categorization, sentiment analysis, question answering, and machine translation, thanks to its use of a transformer architecture.



- **Flan T5 :** FlanT5 is an enhanced version of T5 that has been finetuned in a mixture of tasks. FLAN-T5, developed by Google Research, has been getting a lot of eyes on it as a potential alternative to GPT-3.

LITERATURE REVIEW

- **BERT Score :** In ODQA operations, exact matching is frequently employed as an evaluation metric. However, BERTScore offers a more precise number as a measure of evaluation.

- **News Dataset Generation :** Text, the date, and the paragraph id are fetched from TOI, News-Byte, Print, Scroll, and NDTV and preserved in a JSON format. Then using the ArchivalQA framework news dataset is generated in the Indian context. Generated QA dataset is saved in feather format consisting of columns – Question, Answer, Paragraph, Ans_start, Ans_end.

DATASET STATISTICS

Stat/source	TOI	Economic Times	NDTV	Scroll	News Byte	Total
#ques(after module 2)	251, 879	417, 612	401, 946	400, 070	396, 491	1, 867, 998
#ques(after module 3)	151, 601	478, 973	301, 579	271, 679	269, 328	1, 473, 160
#ques(after module 4)	52, 063	151, 304	106, 321	70, 075	101, 995	**481,758**

DATASET STATISTICS

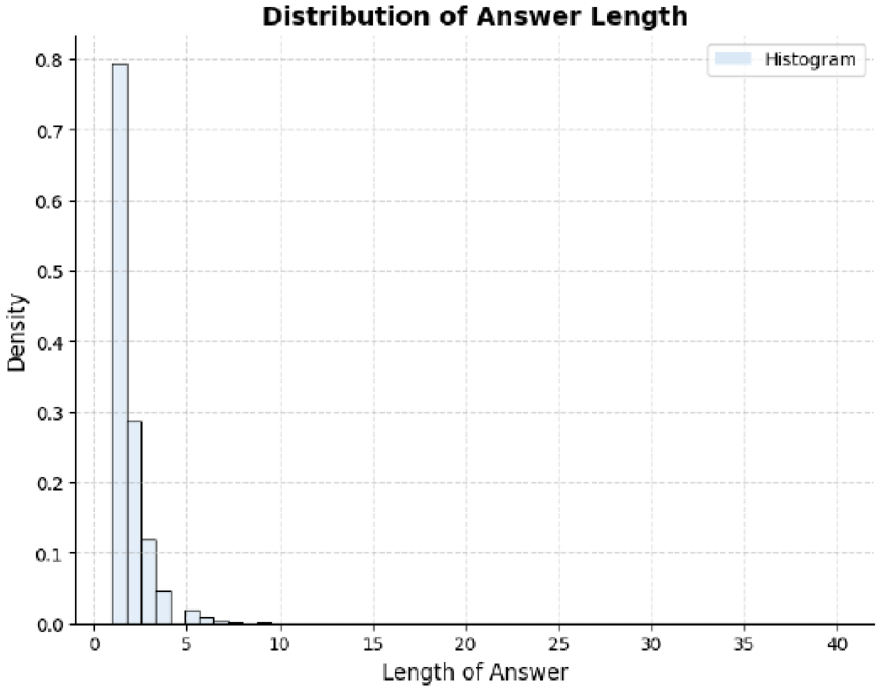


Fig. 2.1 Distribution of Answer Length

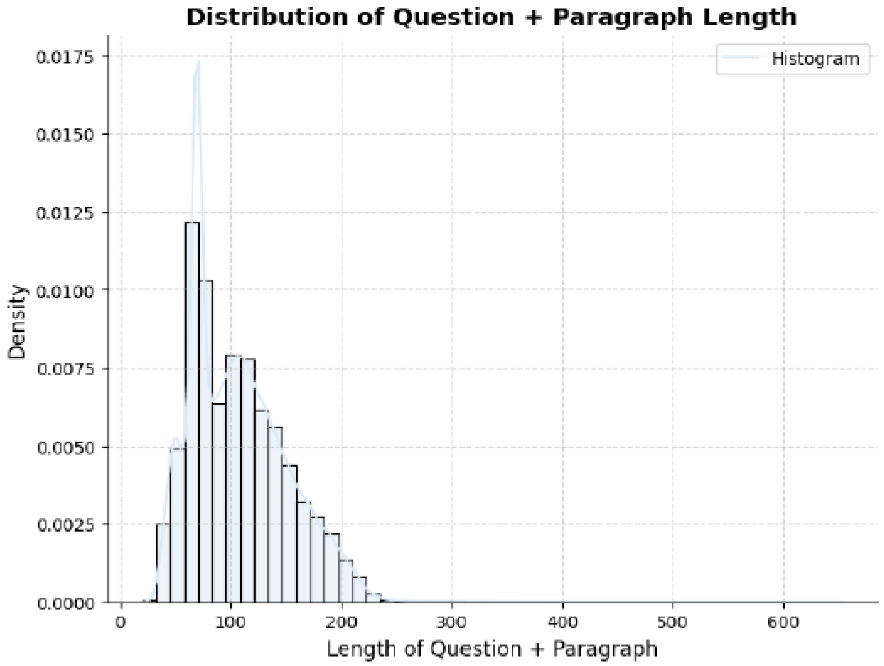


Fig. 2.2 Distribution of Question + Paragraph Length

TIMELINE OF PAPERS

2016

SQUAD

Dataset on which used models are trained

2017

NEWSQA

Previous Popular Dataset on News QA

2019

BERT SERINI

Model used for Evaluation of the Generated Dataset

2020

BERT SCORE

Metric used for Evaluation of the Dataset

2020

T5 TRANSFORMER

Main module in ArchivalQA for Generation of QA...

2022

ARCHIVALQA

Main Framework used to Generate the News QA Dataset

2022

FLAN T5

Enhanced version of T5, finetuned for mixture of tasks and...

CONTRIBUTION

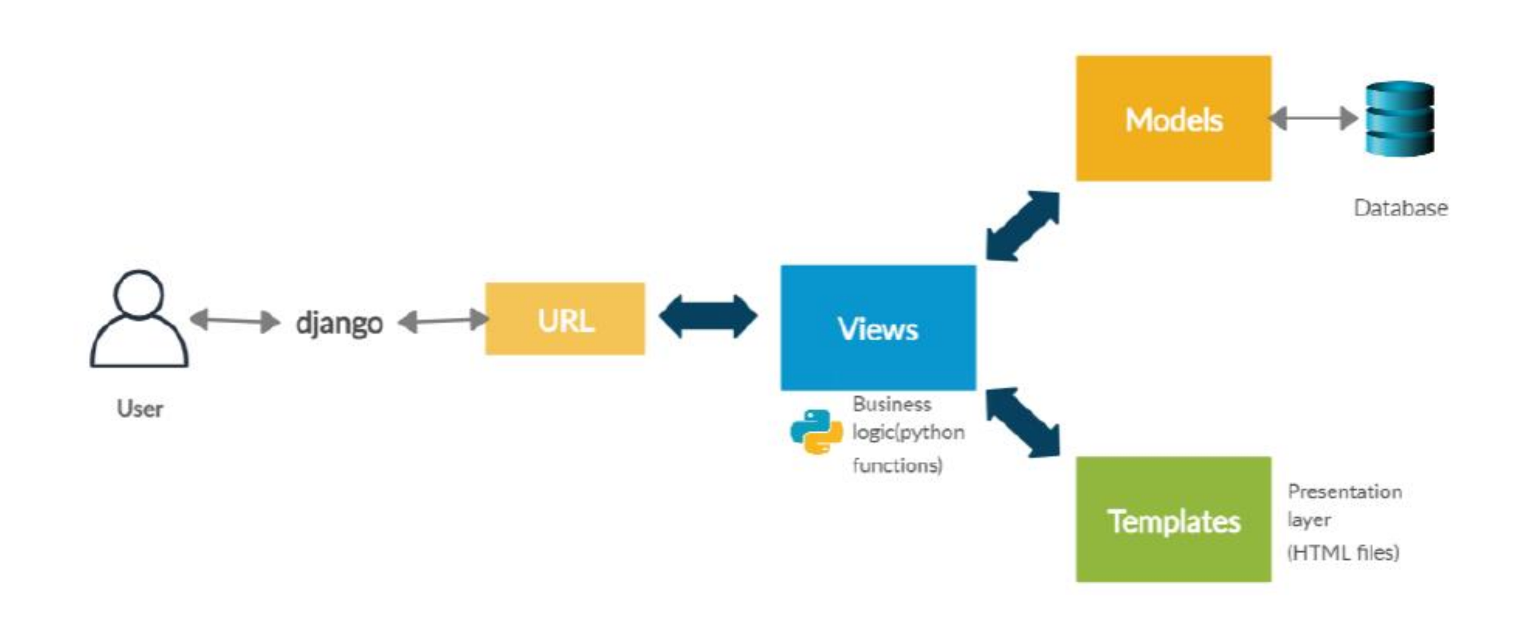
- ❑ Web App based Dataset Verification

- ❑ NLP based Dataset Verification

- ❑ Generative Model NewsQA
    - ❑ QA Model using FlanT5
    - ❑ Context fetching using BM25
    - ❑ Context fetching using KNN
    - ❑ Context fetching using SBERT

- ❑ Teacher Student Model for QA
    - ❑ Attention based Encoder-Decoder Model
    - ❑ Improvement of Attention based Encoder-Decoder Model using Teacher-Student Method

WEB APP BASED DATASET VERIFICATION



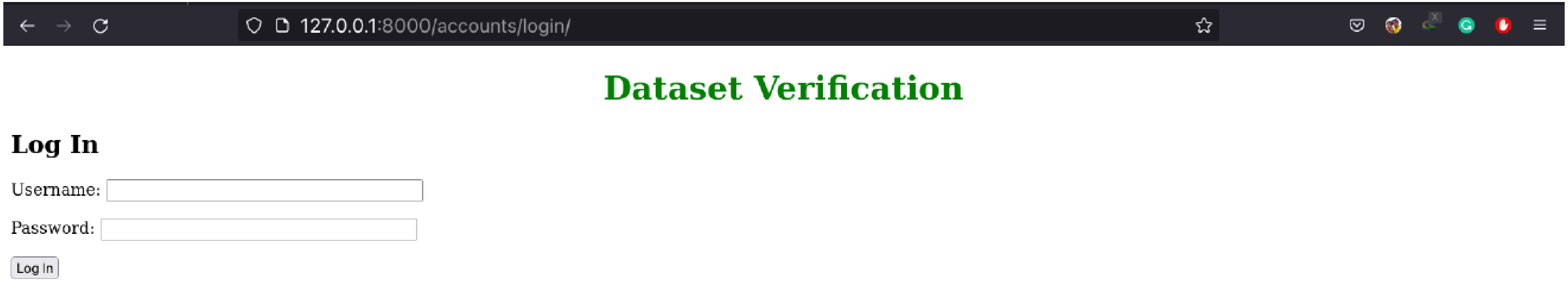
WEB APP BASED DATASET VERIFICATION

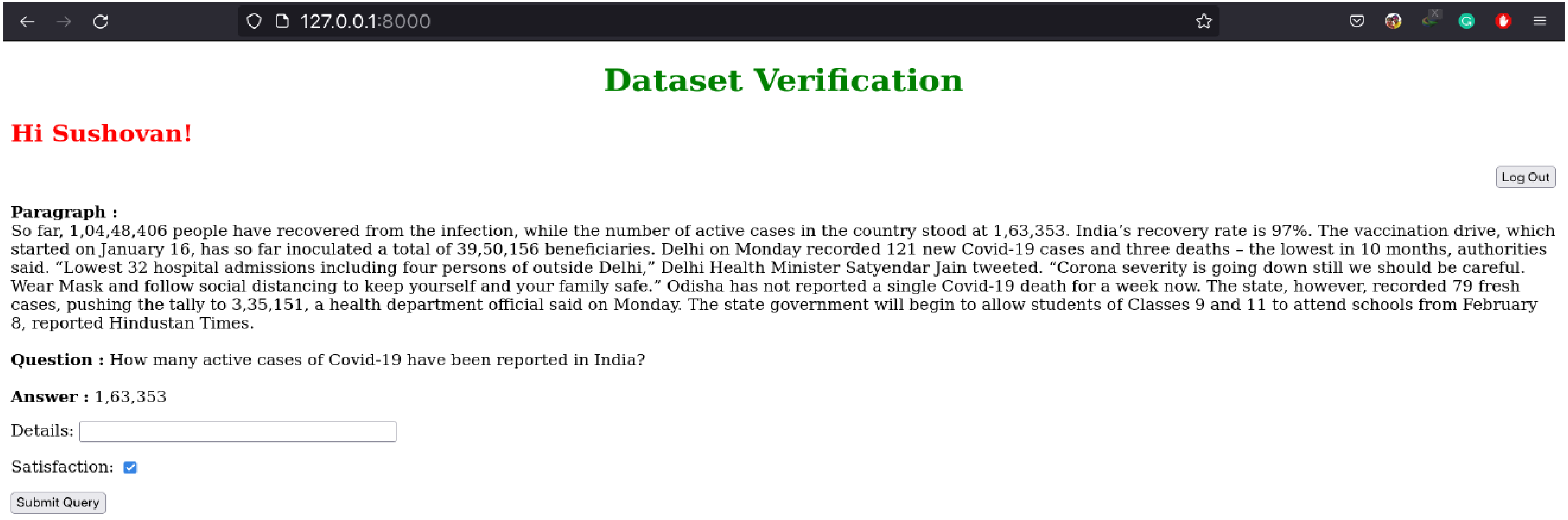
- Our web application, which is based on MVT architecture, chooses an example from the QA dataset that is created.

- The triplet containing **a question, an answer, and a paragraph** serves as the model for this QA pair.

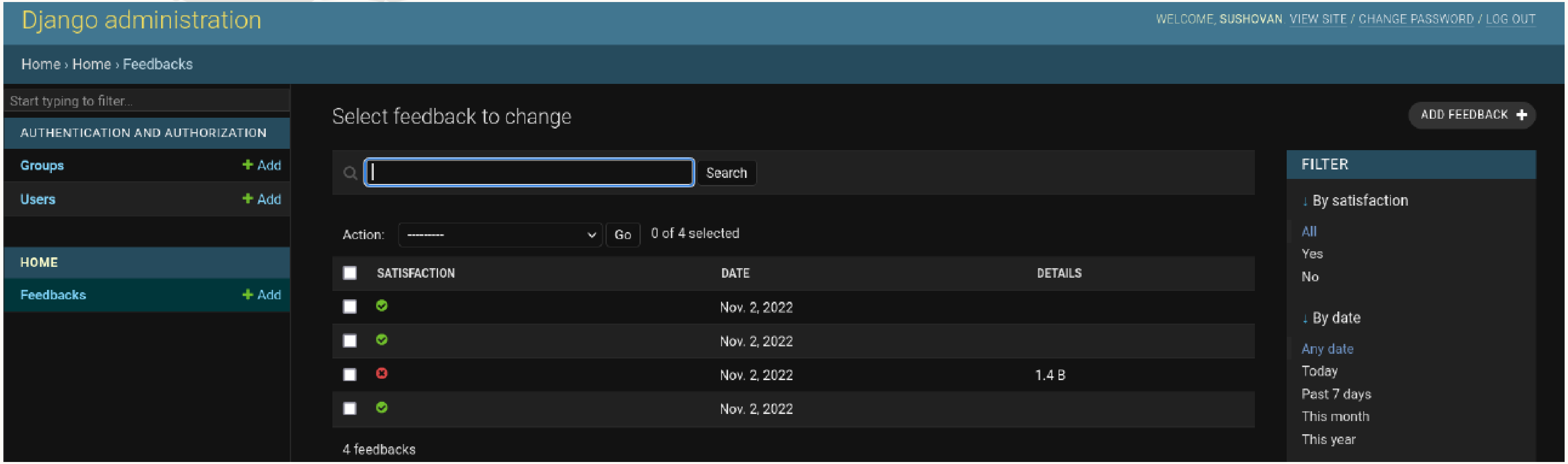
- Now, a person with an account logs in to the website and manually verifies, by reading the paragraph, if the produced response is accurate for the related question.

WEB APP BASED DATASET VERIFICATION





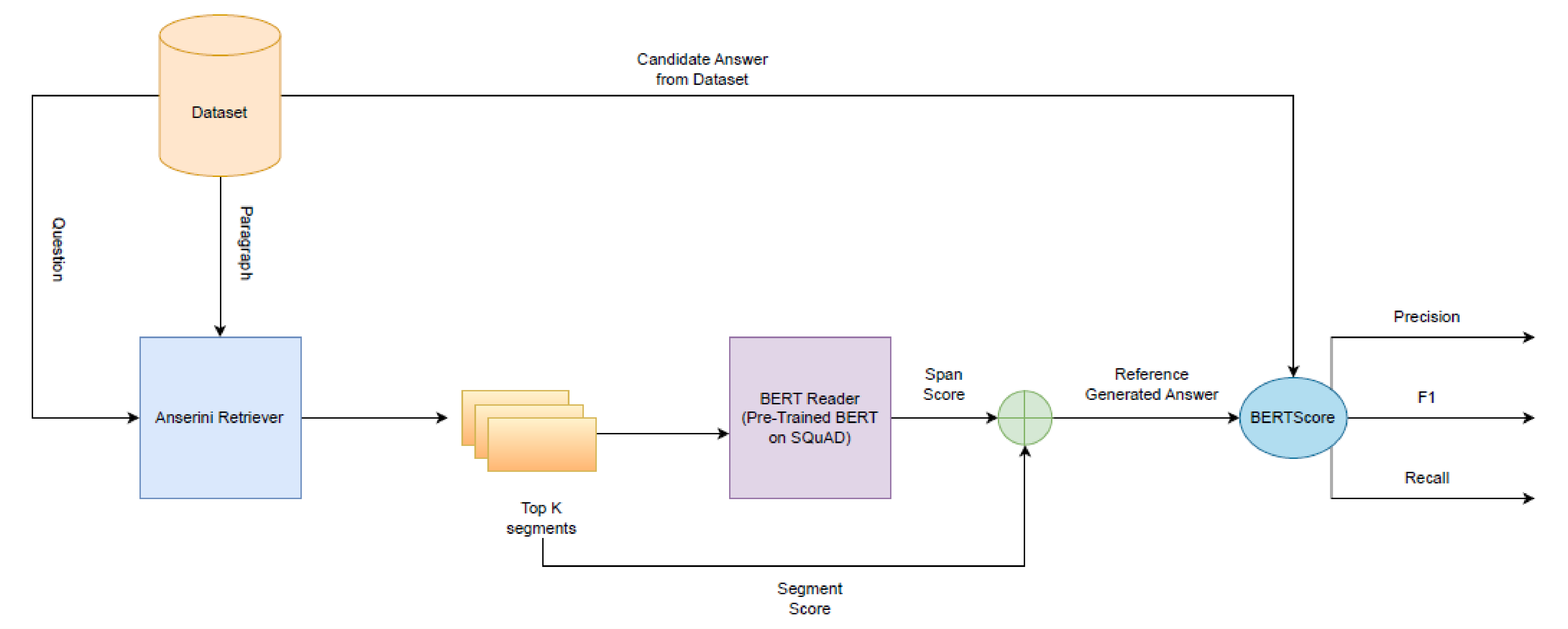
WEB APP BASED DATASET VERIFICATION



NLP BASED DATASET VERIFICATION

- Generate **reference** Answers using pre-trained **BERTSerini**.
- Evaluation of **candidate** answer against **reference** answer using **BERTScore.**

NLP BASED DATASET VERIFICATION

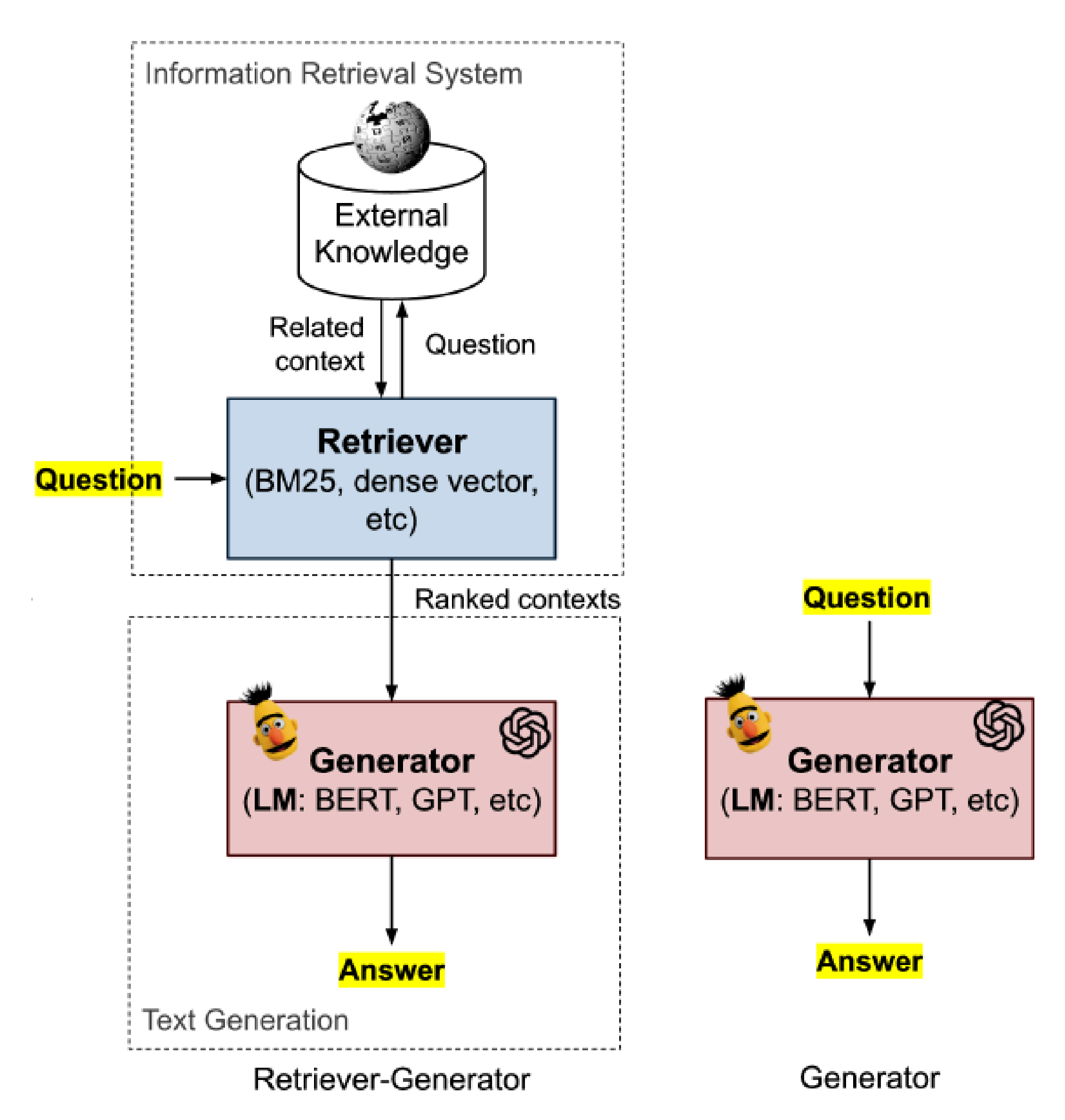


NLP BASED DATASET VERIFICATION : EXPERIMENT AND RESULTS

BERTSerini's BERT Score on our Dataset, which was pre-trained on the SQuAD Dataset

Model	Precision	Recall	F1
BERTSerini	0.521	0.552	0.574

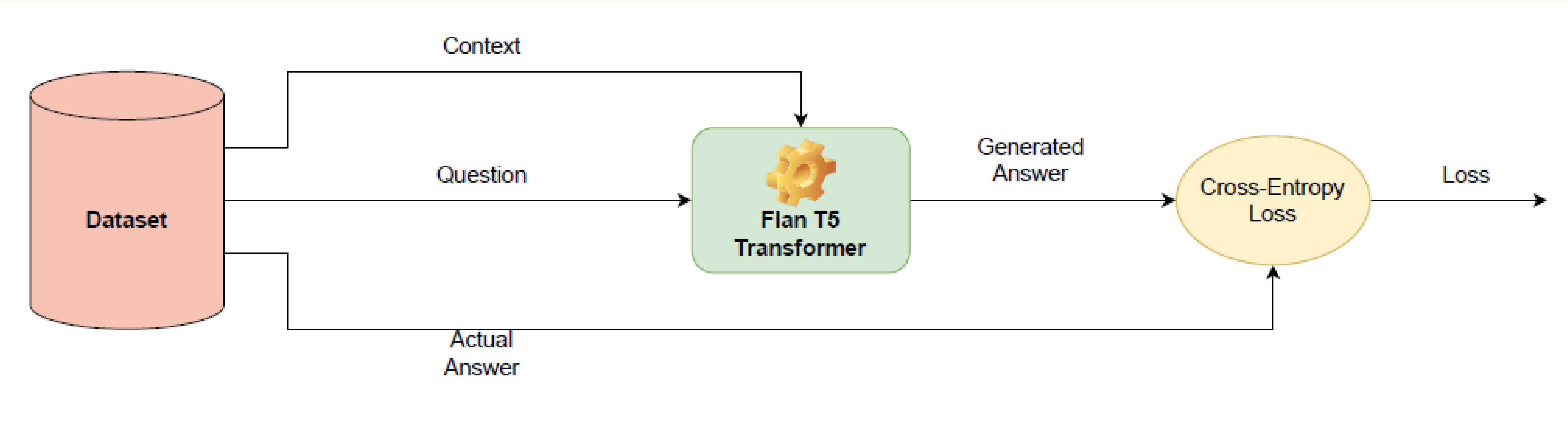
GENERATIVE MODEL FOR NEWSQA



QA MODEL USING FLANT5

- **Question** and **Paragraph** are merged and **tokenized**.
- Tokenized **Answers** served as **Label** to our model.
- **Loss Function : Cross-Entropy** : It is typically applied to the output of the model, which is a probability distribution over the vocabulary for each position in the output sequence.
- **Total Parameters : 247 M**
- Later **METEOR, ROGUE, BLEU** and **BERT** Scores are calculated.

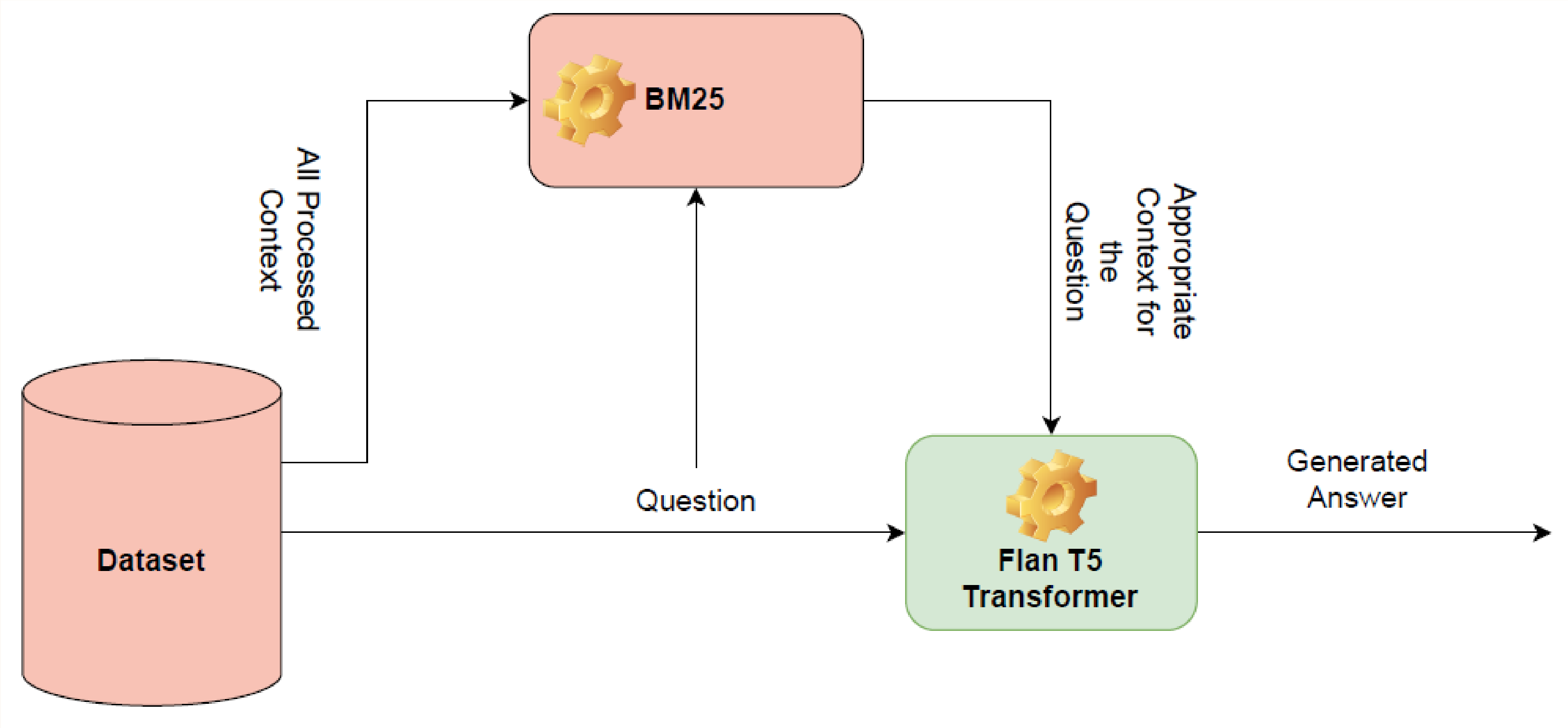
QA MODEL USING FLANT5



CONTEXT FETCHING USING BM25

- ❑ User only provides **Question** to QA System.
- ❑ For generating **answers** we have to fetch **Paragraph** as well.
- ❑ So, we need to modify the **Retriever**.
- ❑ Here we use BM25 as our Retriver. It is based on Tf-Idf vectorization
- ❑ It does not consider the semantic relationship
- ❑ Not suitable for Scalability.
- ❑ Very Old Method

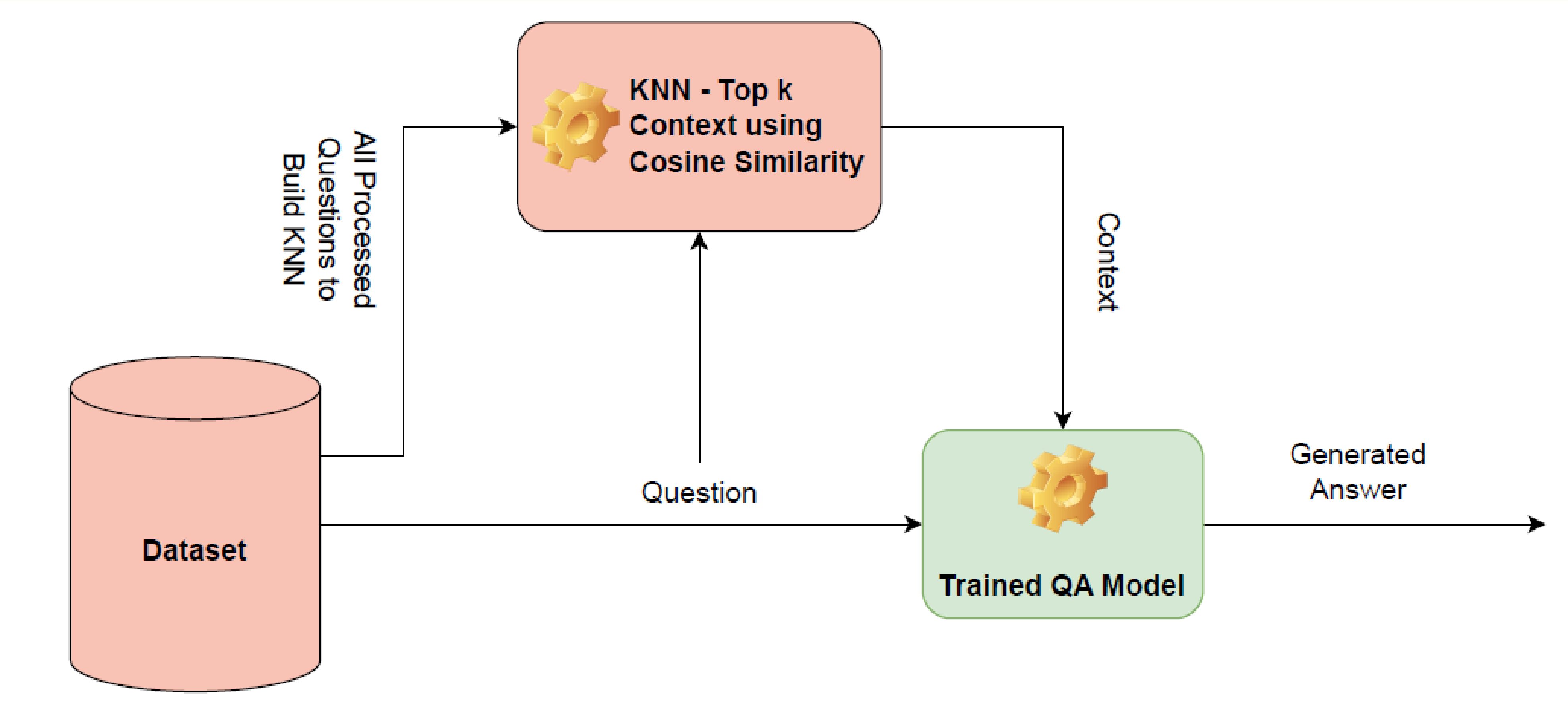
CONTEXT FETCHING USING BM25



CONTEXT FETCHING USING KNN

- Here we are using KNN as Retriver.
- Questions are vectorized using TF-IDF for KNN. Using KNN top k nearest questions are fetched based on cosine distance, and their corresponding paragraph is returned.
- It works fast as , There are no trainable parameters in the traditional (KNN).
- It does not consider the semantic relationship as based on TF-IDF
- Not suitable for Scalability.

CONTEXT FETCHING USING KNN



CONTEXT FETCHING USING SBERT

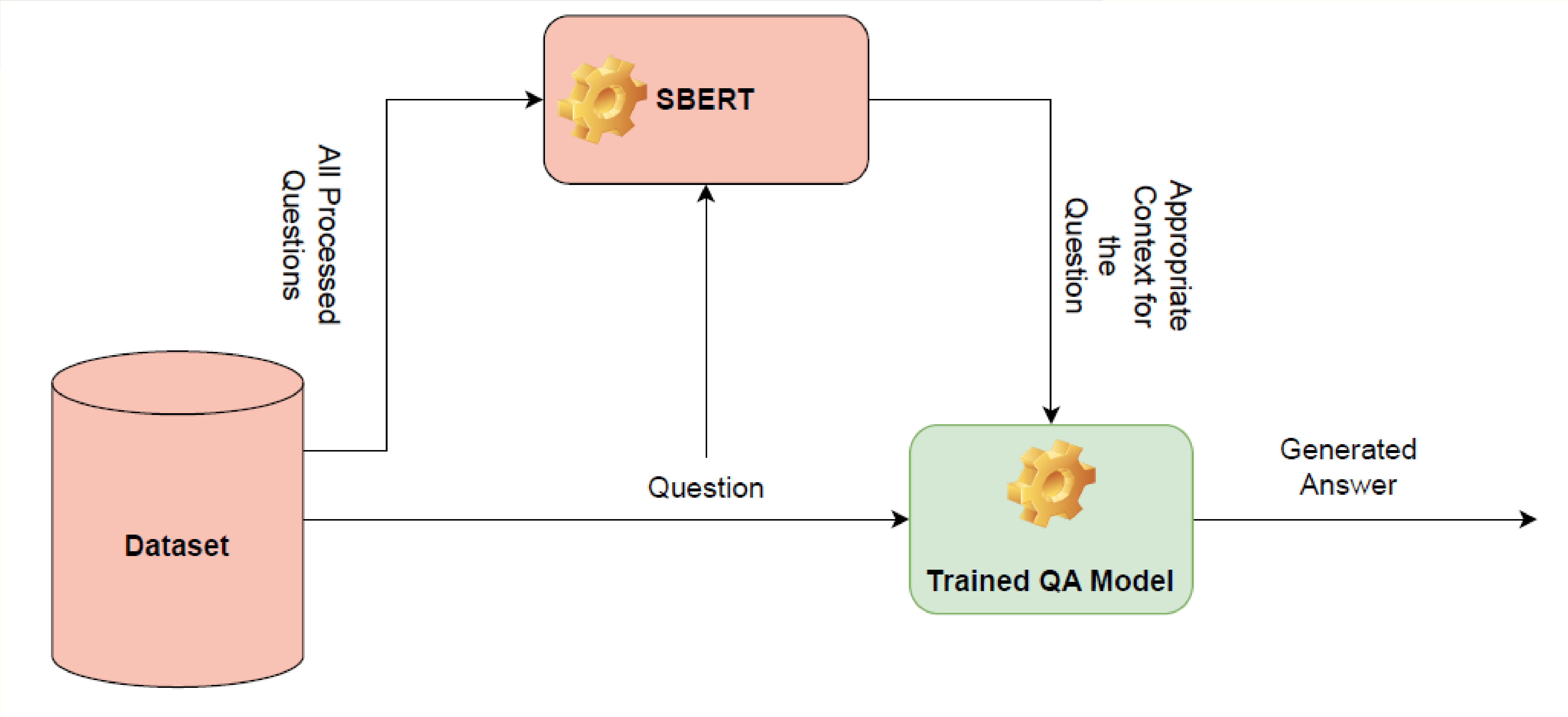
- State-of-the-art Pre-Trained **SBERT** is used as **Retriever**.

- It can recognize the **semantic** connections between terms and creates a **concise** yet **information rich** vector for sentences, accelerating retrieval calculation.

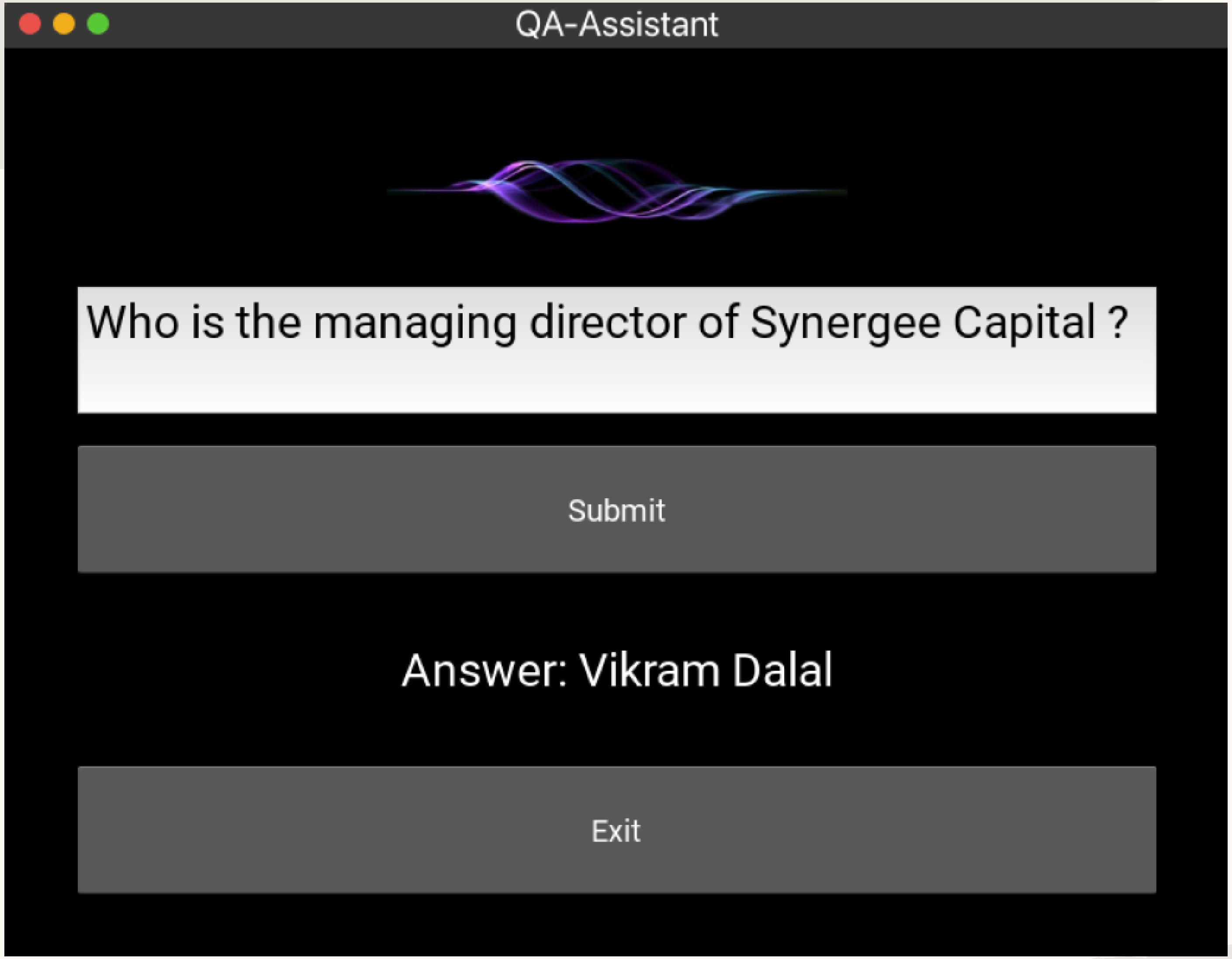
- The pre-trained **all-MiniLM-L6-v2** model can encode 14200 sentences per second on a V100 GPU.

- Suitable for **Scalability**.

CONTEXT FETCHING USING SBERT



GUI FOR QA MODEL

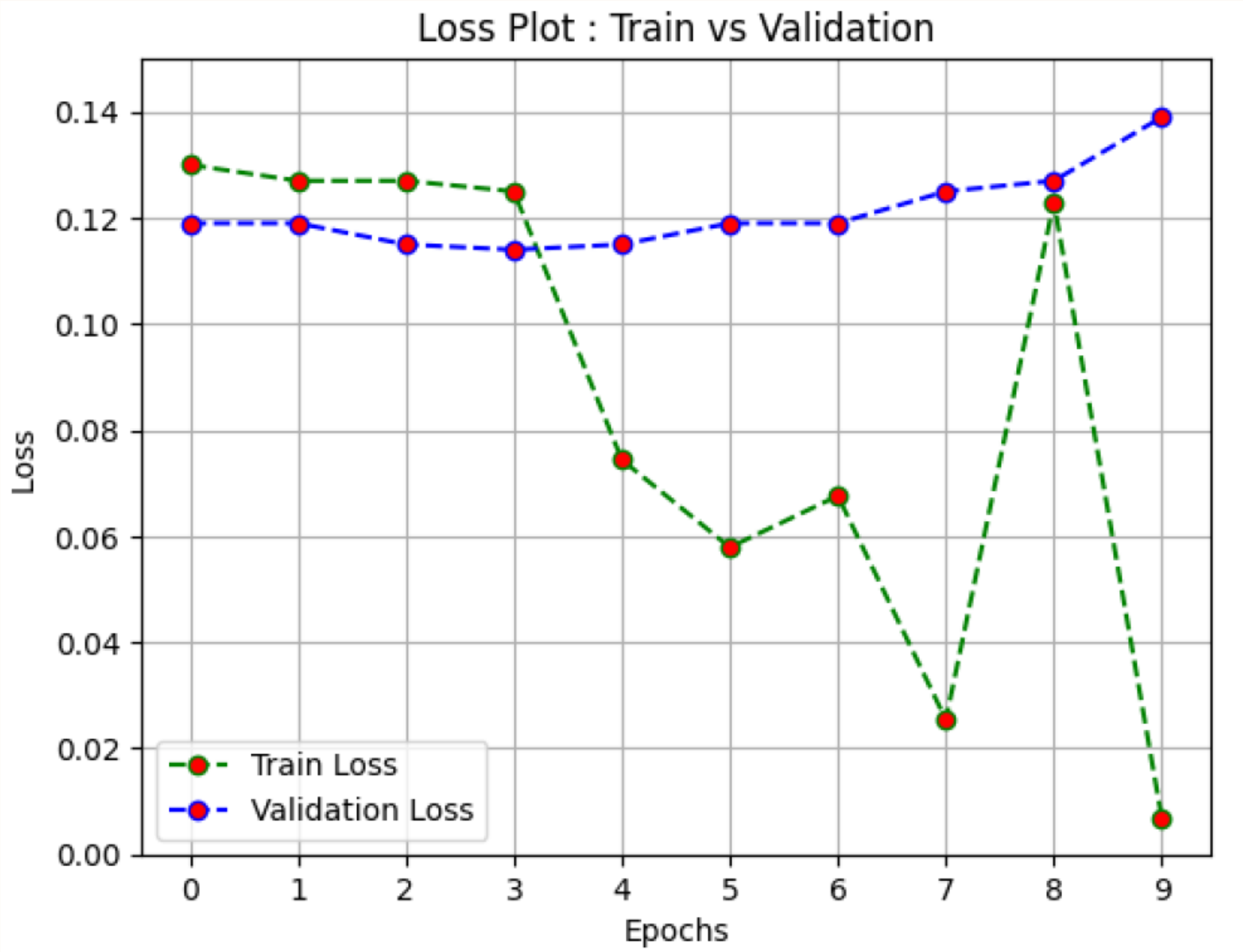


QA MODEL : EXPERIMENT AND RESULTS

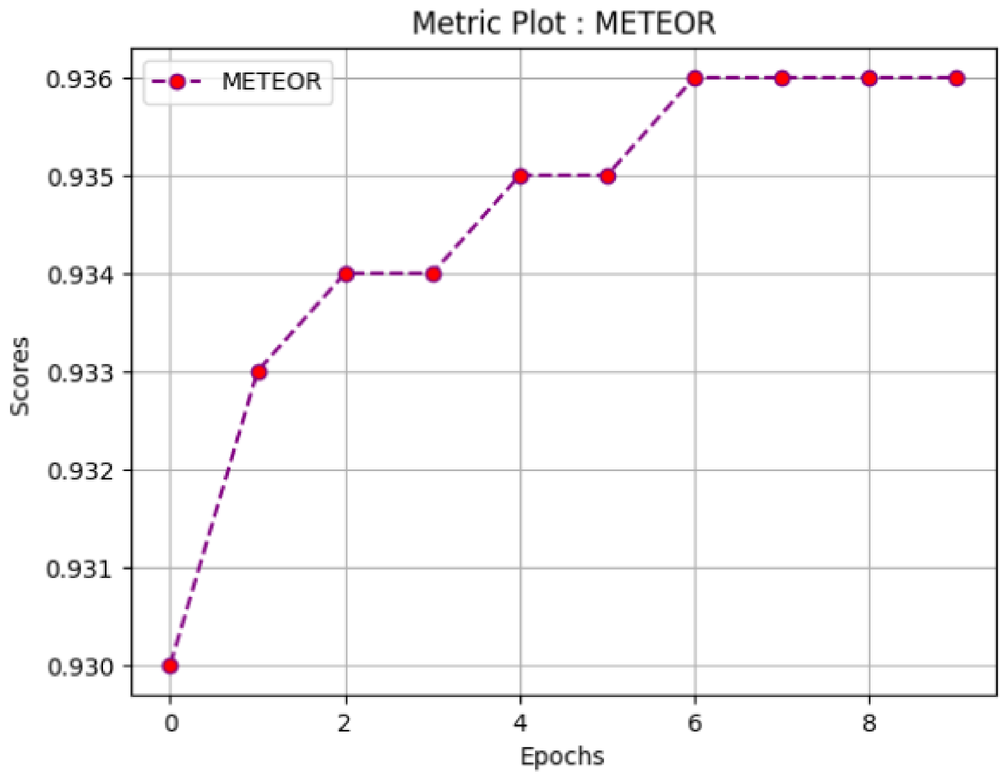
Results on Test Batch Data

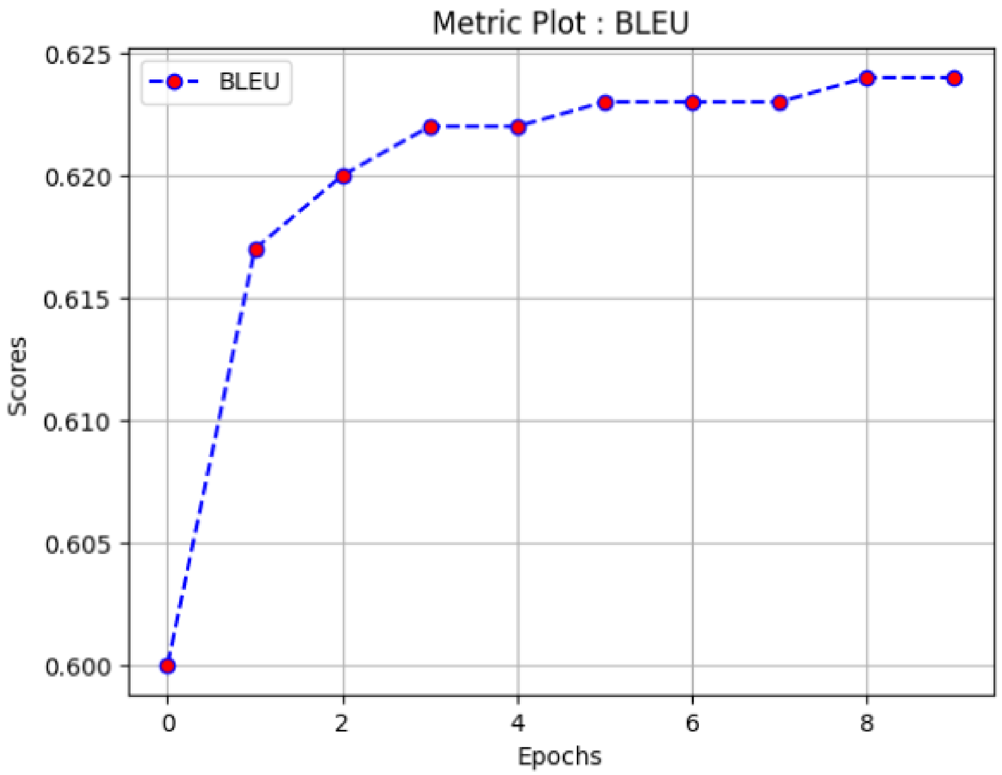
Model	Loss	BLEU	METEOR	ROUGE
QA-Model	0.144	0.623	0.934	0.934

QA MODEL : EXPERIMENT AND RESULTS



QA MODEL : EXPERIMENT AND RESULTS





QA MODEL : EXPERIMENT AND RESULTS

BLEU, METEOR and ROUGE Scores on Full Dataset

Model	BLEU	METEOR	ROUGE
QA-Model	0.648	0.966	0.965

BERT Scores on Full Dataset

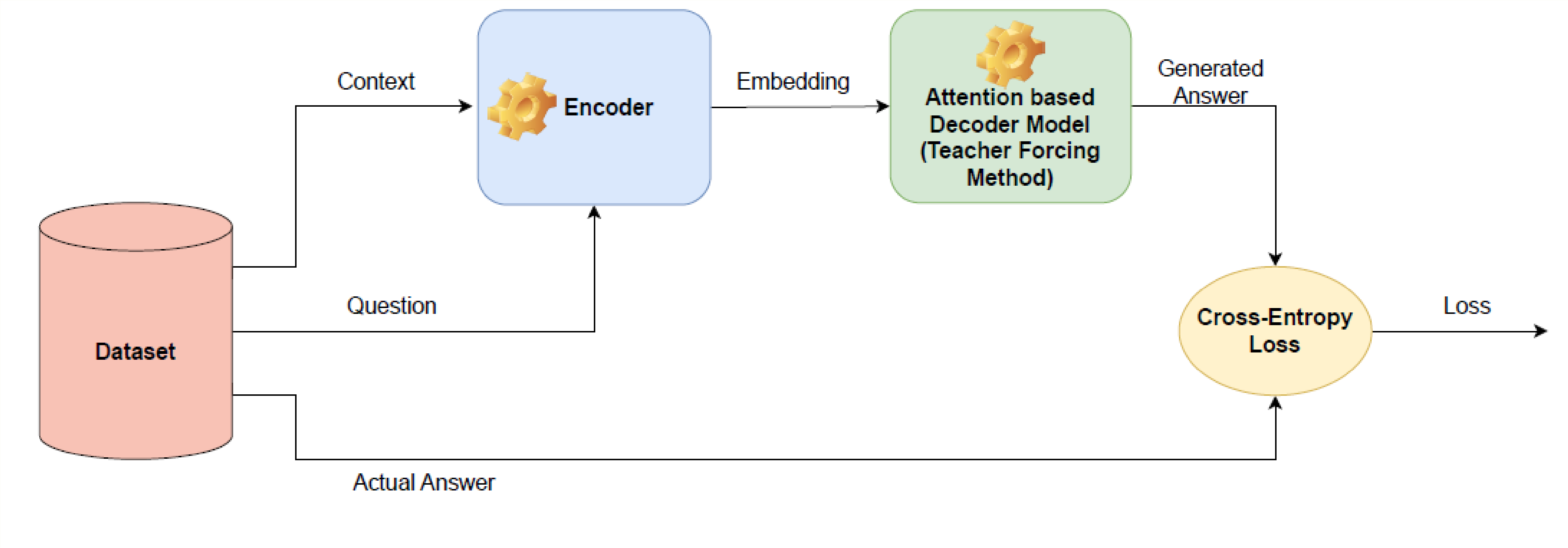
Model	PRECISION	RECALL	F1
QA-Model	0.966	0.964	0.964

TEACHER STUDENT METHOD FOR QA

ATTENTION BASED ENCODER-DECODER MODEL FOR QA

- **Low parameter, Less Complex** model designed for QA task.
- **Bidirectional GRU** (Gated Recurrent Unit) is used as the main building block for both the encoder and decoder.
- Model is trained using **Teacher-Forcing** Method.
- **Cross-Entropy** is used as Loss Function.
- **Total Parameters : 143 M**
- This Model is not so powerful for QA Task.

ATTENTION BASED ENCODER-DECODER MODEL FOR QA



ATTENTION BASED ENCODER-DECODER MODEL FOR QA

Cross-Entropy Loss on Batch Test Data

Model	Cross-Entropy Loss
Vanilla Model	4.585

Cross-Entropy Loss on Full Dataset

Model	Cross-Entropy Loss
Vanilla Model	4.654

ATTENTION BASED ENCODER-DECODER MODEL FOR QA

IMPROVEMENT USING TEACHER STUDENT METHOD

- Here we want to improve previous model using Teacher Student Mechanism.

- Previously trained **FlanT5 based QA Model** is used as **Teacher Model**.

- **Attention based Encoder-Decoder** model is used as **Student model**.

- Here Student Model try to learn better **Encoder Embedding** from Teacher Model in order generate better Decoder Output.

- Along with Cross-Entropy, KL-Divergence loss is also used between both encoders' outputs.

ATTENTION BASED ENCODER-DECODER MODEL FOR QA

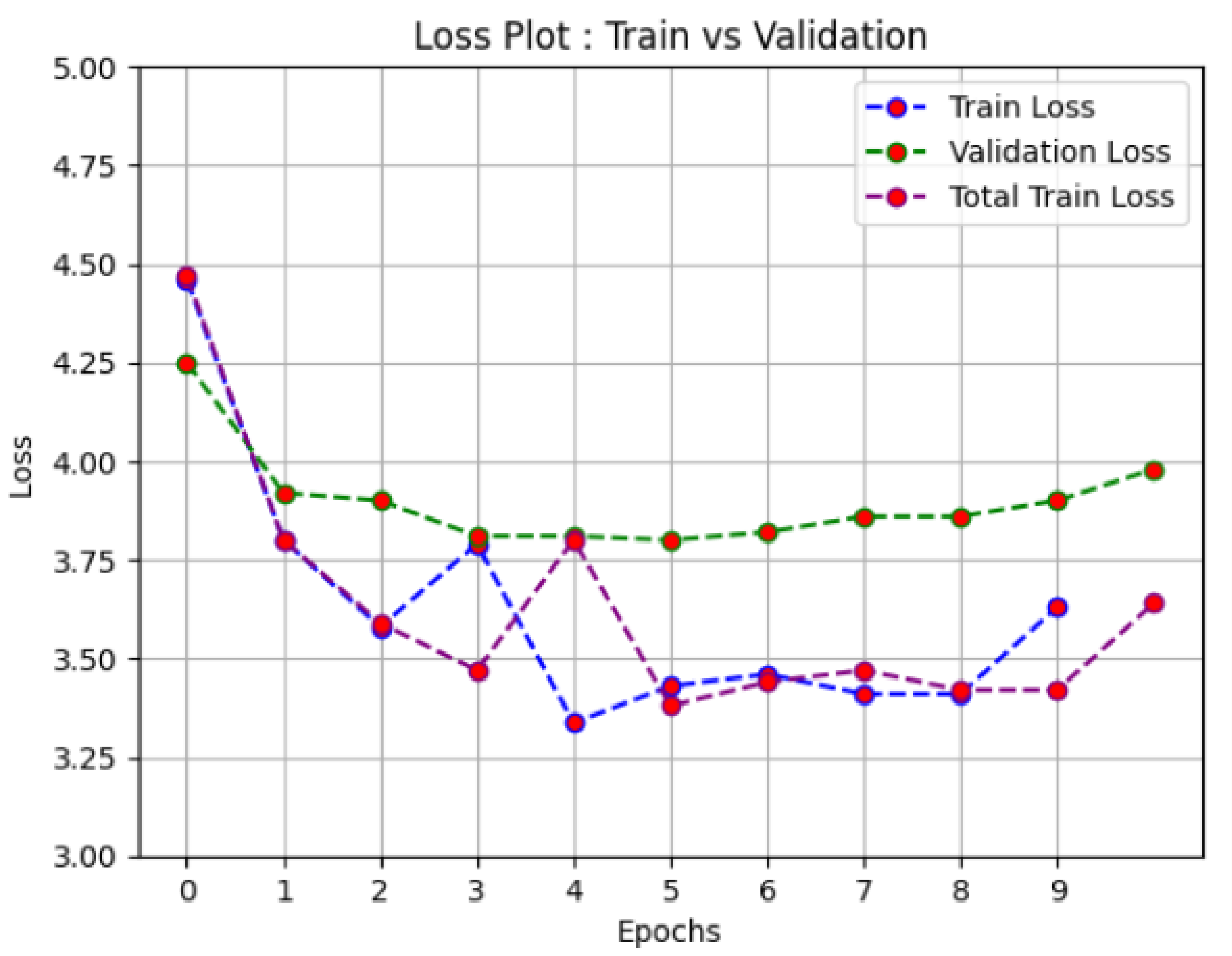
Cross-Entropy Loss on Batch Test Data

Model	Encoder Loss	Cross-Entropy Loss	Total Loss
Model (After TS)	0.0071	3.797	3.804

Cross-Entropy Loss on Full Dataset

Model	Encoder Loss	Cross-Entropy Loss	Total Loss
Model (After TS)	0.0073	3.862	3.883

ATTENTION BASED ENCODER-DECODER MODEL FOR QA



CONCLUSION & FUTURE WORK

- Using this Dataset Generation Framework QA Dataset can be Extended in any domain.

- The current generative model is limited to the dataset it was trained on and lacks access to global knowledge; incorporating state-of-the-art APIs provided by various organizations can enhance its knowledge base.

- Furthermore, the birth of large language models(LLMs) with billions of parameters opens up new possibilities. Exploring these larger language models, such as ChatGPT or future iterations, can lead to more accurate QA results.

- The attention-based encoder-decoder model was less complex; further optimization, fine-tuning, and a more complex base model could yield even better results.

THANK YOU