NIC

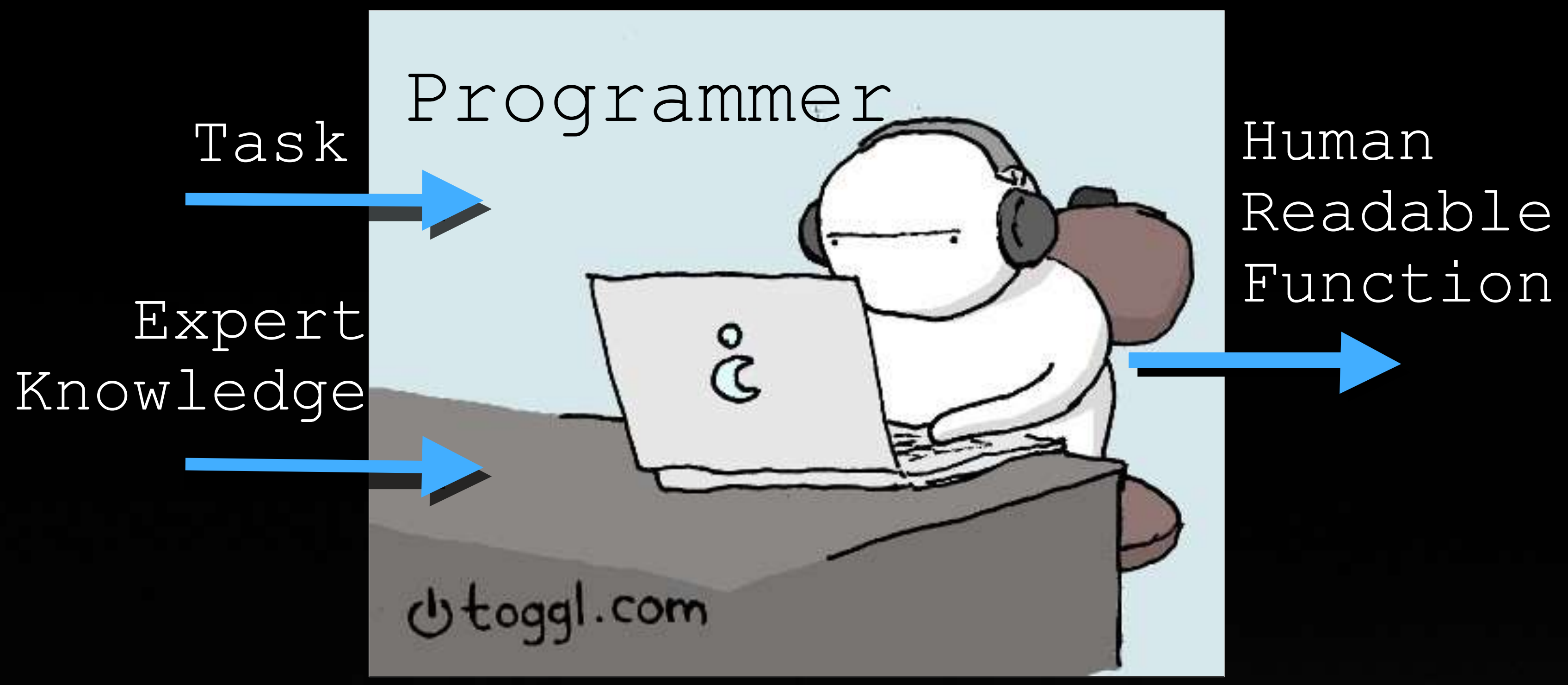
NIC AI HACKATHON 2022

AUTOMATED DOCUMENT Q & A EXTRACTION MODEL

A NEW WAY TO BUILD SOFTWARE

Traditional Programming vs Machine Learning

SOFTWARE 1.0: Traditional Programming



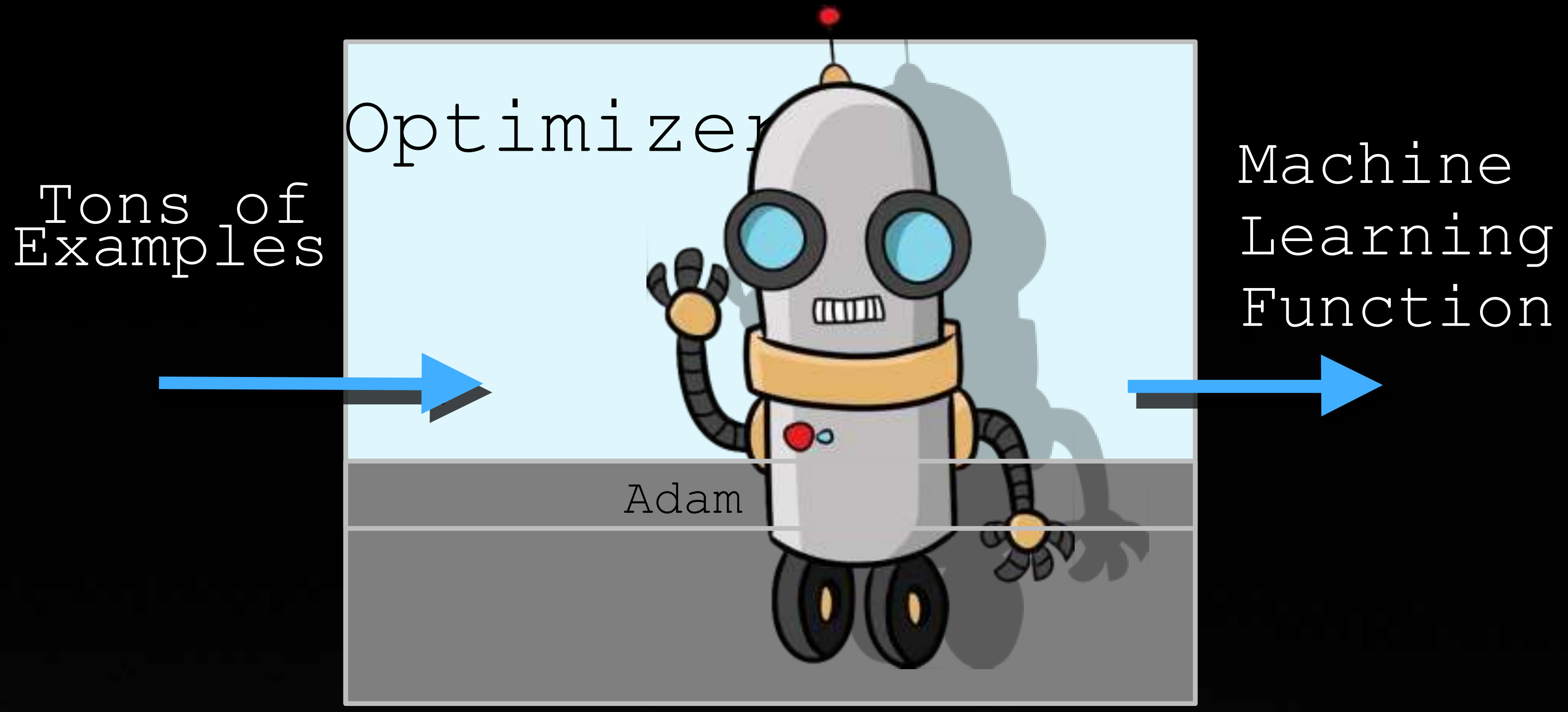
Task

Expert Knowledge

Programmer

Human Readable Function

SOFTWARE 2.0: Machine Learning

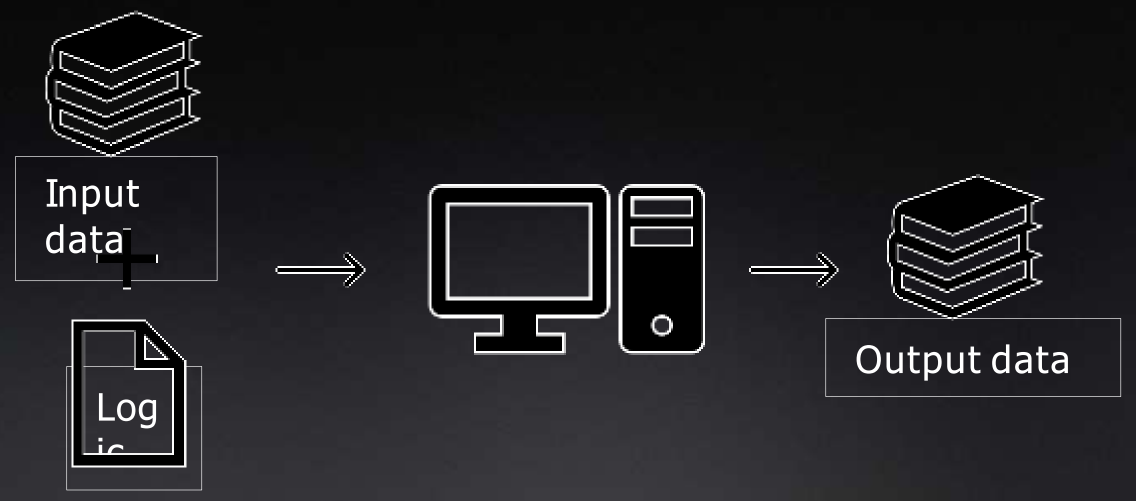


Tons of Examples

Optimizer

Adam

Machine Learning Function



Input data

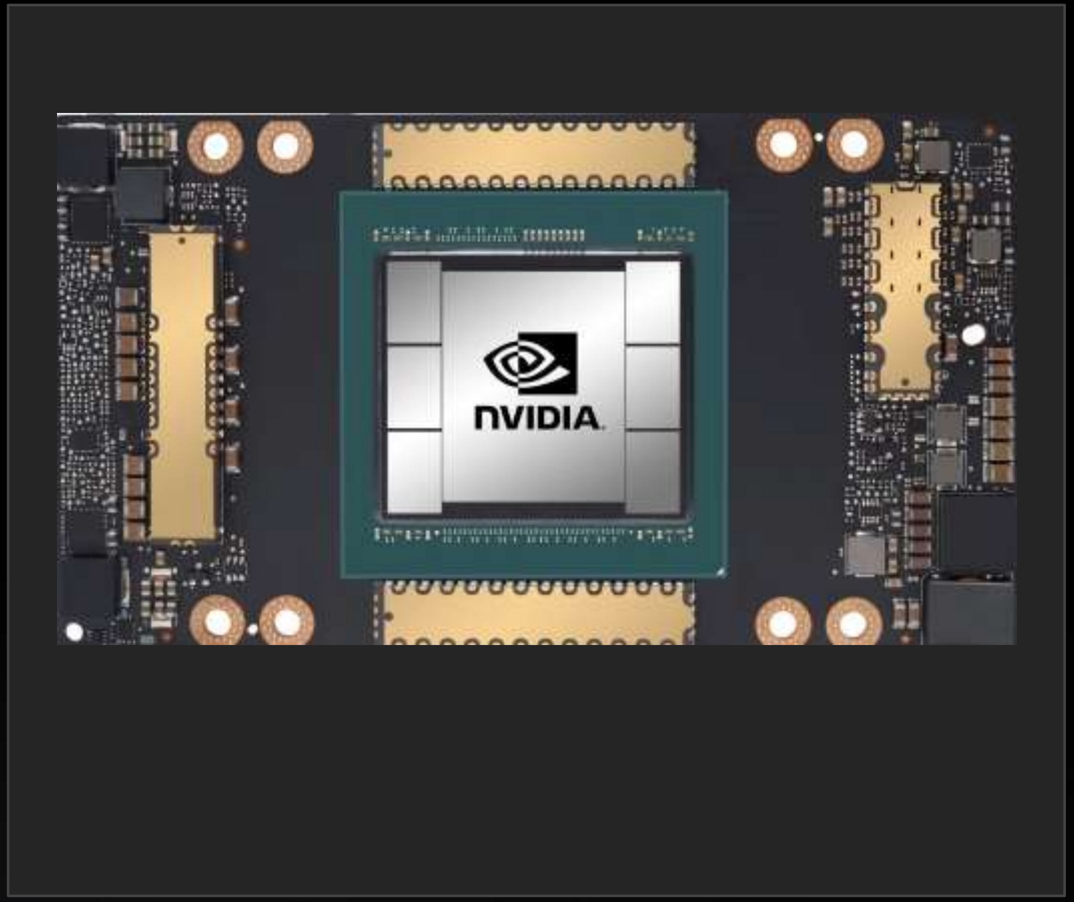
Logic

Output data

You need three main ingredients (and some

LARGE QUANTITIES OF DATA

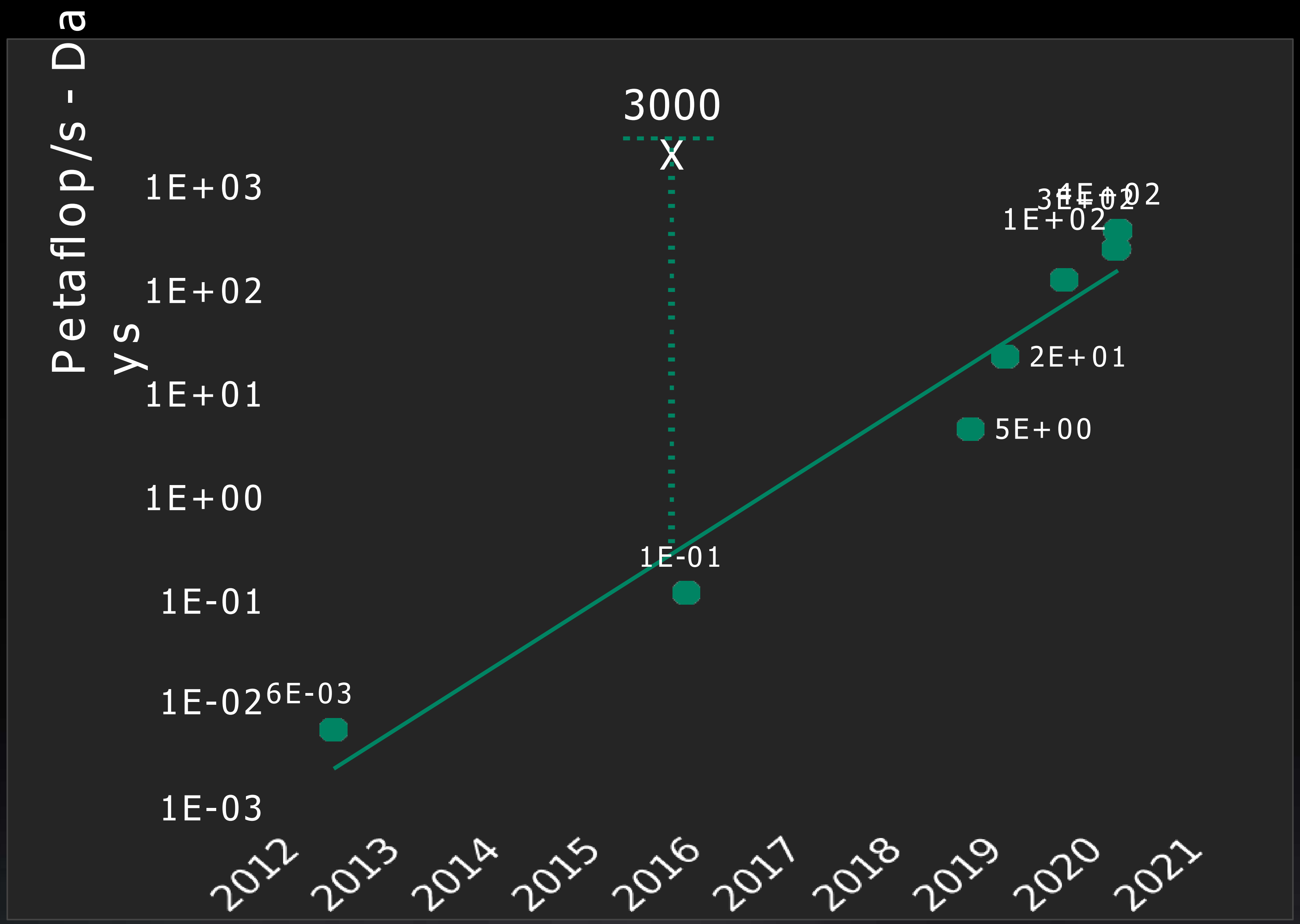
ML FRAMEWORK



GPU ACCELERATOR

CHALLENGES: ACCELERATING BIG AND SMALL

AI Advances Demand Exponentially Higher Compute



3000X Higher Compute Required to Train Largest Models Since Volta

Source: OpenAI, NVIDIA

Problem Statement

Automated Document Q&A Extraction Model

- Create an AI Model for Document Q&A Services on Government Dataset.
- Government Documents like Acts/ Policies/Rules/ Guidelines/ Notifications/ Frequenly Asked Questions are increasingly accessed by Citizens in day to day life for Ease of Doing Business with Government & Other entities.
- In light of Ease of Doing Business, we need to ease out this process using Intelligence Augmentation through AI.
- Create a Proof of Concept AI Model and then fine tuning the model to increase accuracy

INTRODUCTION

- NIC (National Informatics Centre) conducted an AI (Artificial Intelligence) based hackathon where each state formed a team and prepared a QA (Question - Answering) model.

- The data was fetched from all portfolios and departments of the Central and State Governments, for example External Affairs, Education, Technology, etc.

- The best models were identified based on the accuracy of models trained.

- The purpose of this hackathon was to improve the use of AI in government organisations and provide an interactive QA based AI assistant for common public usage.

INTRODUCTION

- AI Hackathon 2022 was divided in 3 stages. Bootcamps were conducted at various stages to understand the basic concepts of AI, NLP, transformer, data tagging, model deployment strategies.

PHASE I :

- Data preparation: Collecting data from various ministries from acts/rules/ regulations, notifications and FAQs. Around 8000 question answer pair were prepared in SQUAD 2.0 . Python scripts were written to convert prepared Q A dataset in json format.
- Model Selection: Various online Q A models such as Roberta , Bert, Distlled Bert , XLm etc were tried and tested on QA dataset. Roberta_x0002_base model was finalized based on evaluation parameters.

INTRODUCTION

PHASE II :

- Model Fine-Tuning: Roberta-base model was fine-tuned using various combinations of hyper parameters and optimized for inference. The
- Pretrained model was trained on the dataset prepared. Various python scripts were written for fine-tuning the model to improve the accuracy of the model. AI libraries like transformer, pyTorch were used to implement functionality. Roberta-base was tested for accuracy and F1 score on the provided test set by AI HQ team and **NIC Sikkim stood 5th at the national evaluation.**

INTRODUCTION

PHASE III :

- Deployment Stage: The model was deployed on triton server . The deployed model was tested on evaluation set provided by AI HQ team.

PHASE I

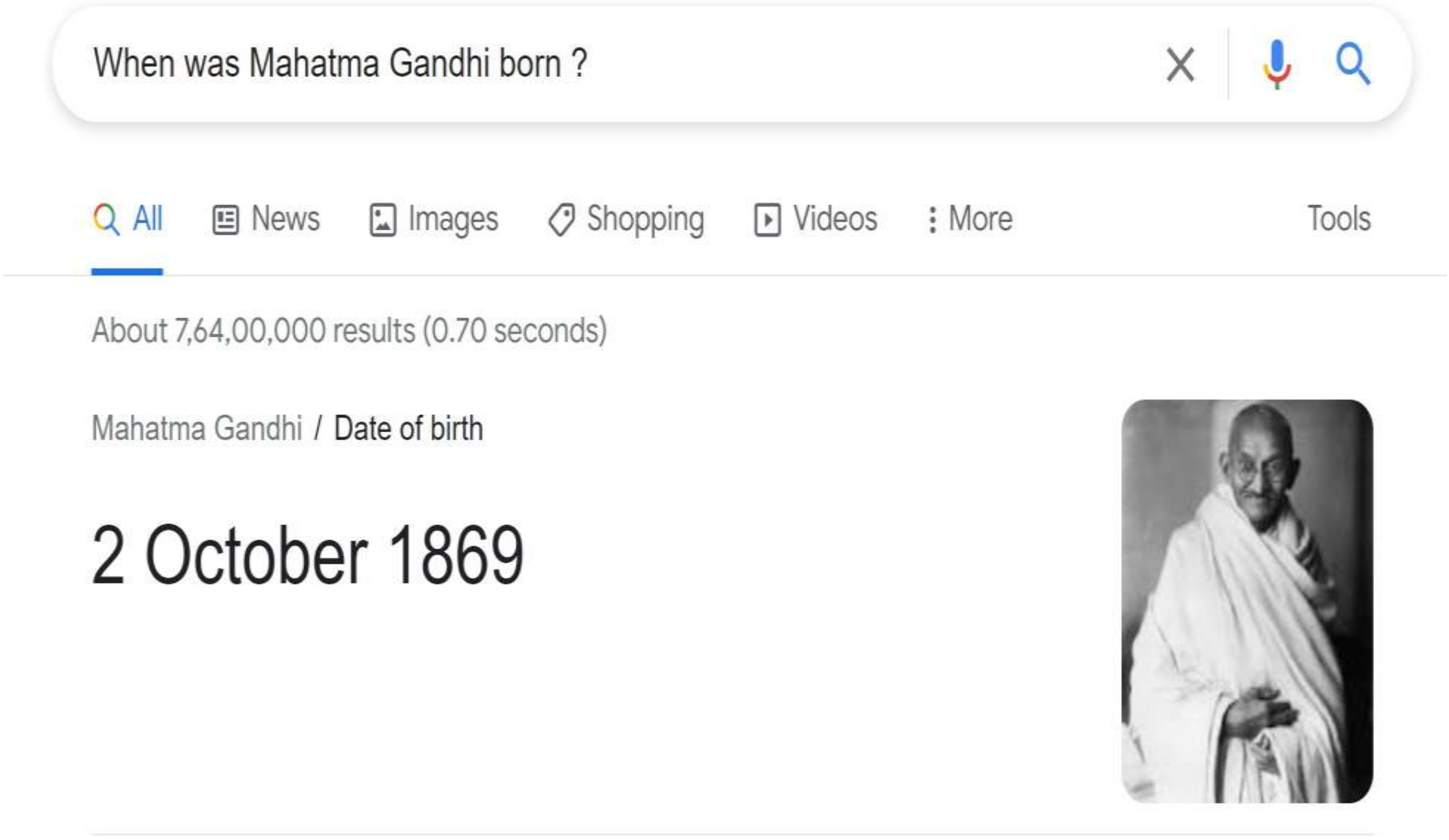
DATASET PREPARATION

WHAT IS QNA SYSTEM ?

Explanation

Question Answering(QA) system is a system that gives appropriate answers to questions expressed in natural languages such as English, Hindi, and so on.

For example, suppose a user asks " *When was Mahatma Gandhi Born?"* In this case, the question answering system is expected to return "2 October 1869".



When was Mahatma Gandhi born ?

About 7,64,00,000 results (0.70 seconds)

Mahatma Gandhi / Date of birth

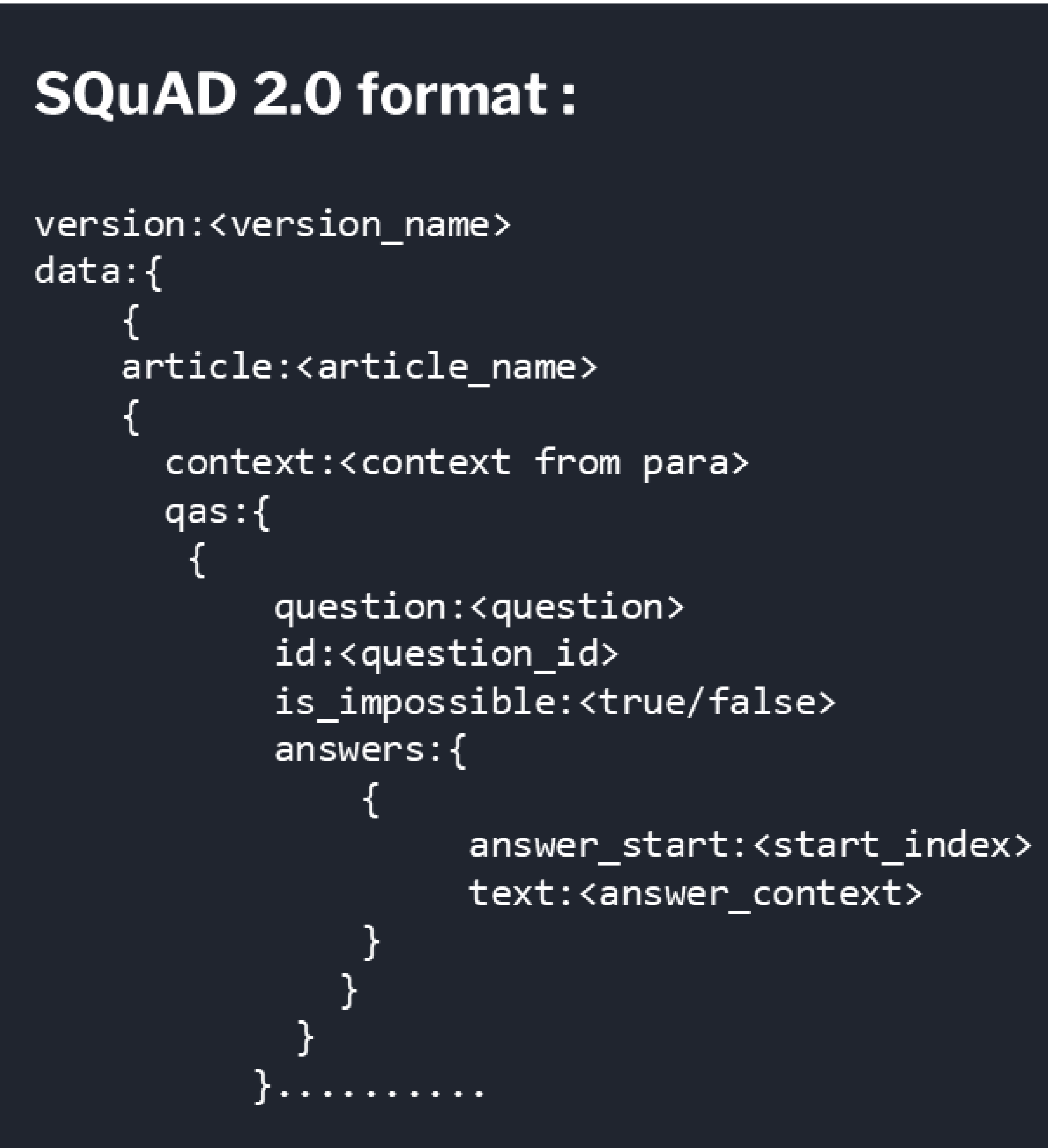
2 October 1869

WHAT IS QNA SYSTEM ?

Government Samples

Questions	Answers
What is the Full Form of BOOT?	Build Own Operate and Transfer
What is date of National Sports Day?	August 29
When was the Pradhan Mantri Krishi Sinchayee Yojana (PMKSY) Launched ?	During 2015-2016
What is the full form of HKKP?	Har Khet Ko Pani (HKKP)
Who does approve the Detailed Project Report (DPR) for the proposal of water bodies ?	State Technical Advisory Committee (TAC)

DATASET FORMAT SQUAD 2.0



SQuAD 2.0 format :

```
version:<version_name>
data:{
    {
    article:<article_name>
    {
      context:<context from para>
      qas:{
       {
           question:<question>
           id:<question_id>
           is_impossible:<true/false>
           answers:{
               {
                    answer_start:<start_index>
                    text:<answer_context>
               }
              }
            }
          }..........
```

- ✓ Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles.
- ✓ The answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable
- ✓ To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.
- ✓ This dataset was chosen due to its simple yet powerful format and easy trainability of data

DATA SOURCES

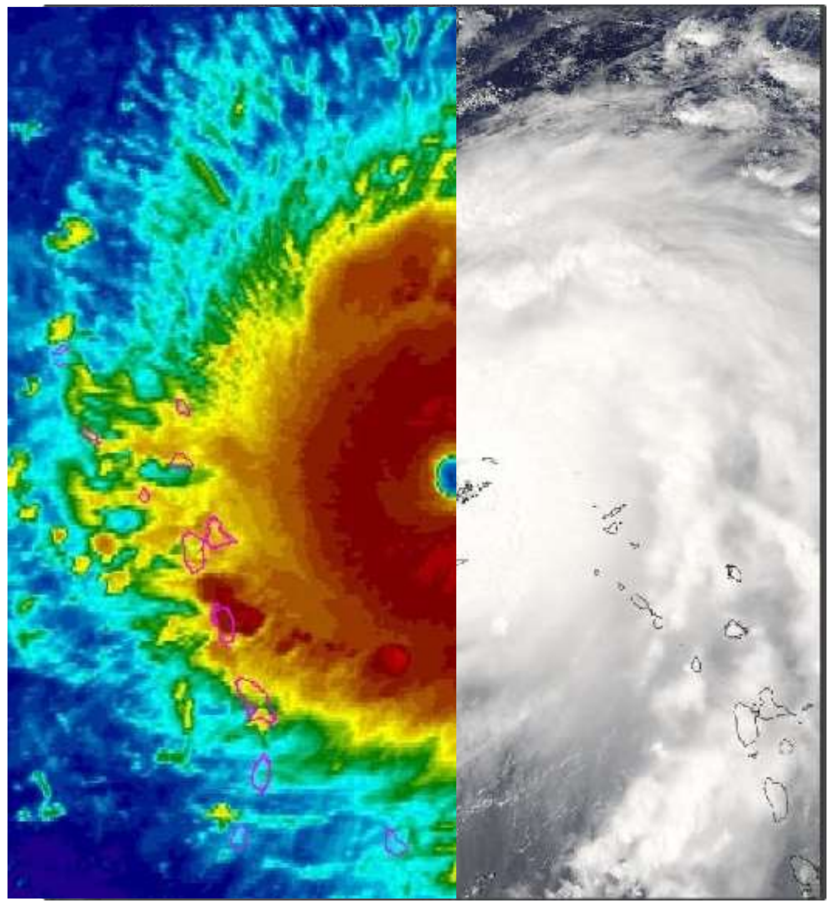
All data for training, testing and validation of the AI model was taken from respective government departmental websites

1 Acts and Rules

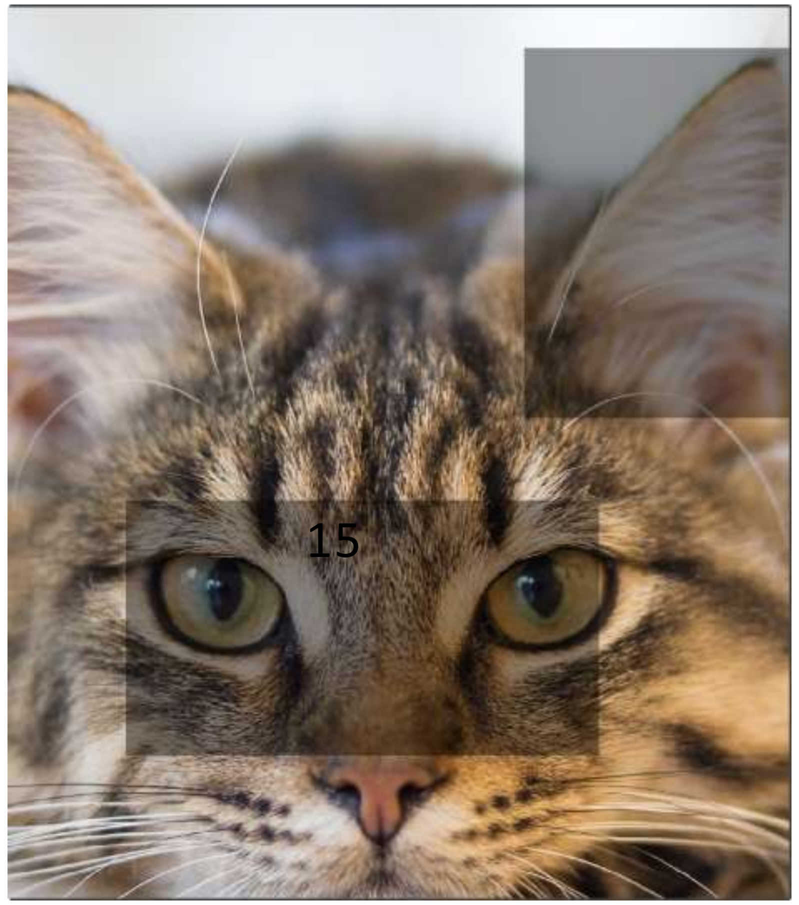
2 Notifications

3 Frequently Asked Questions

LABELLING LARGE QUANTITIES OF DATA

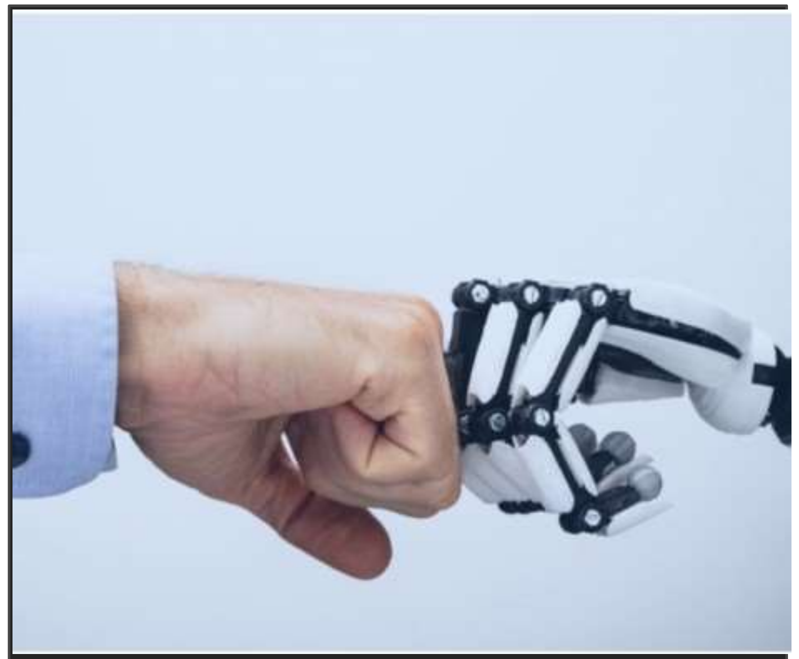


Using one data source as the label for another



Predicting input B from input A

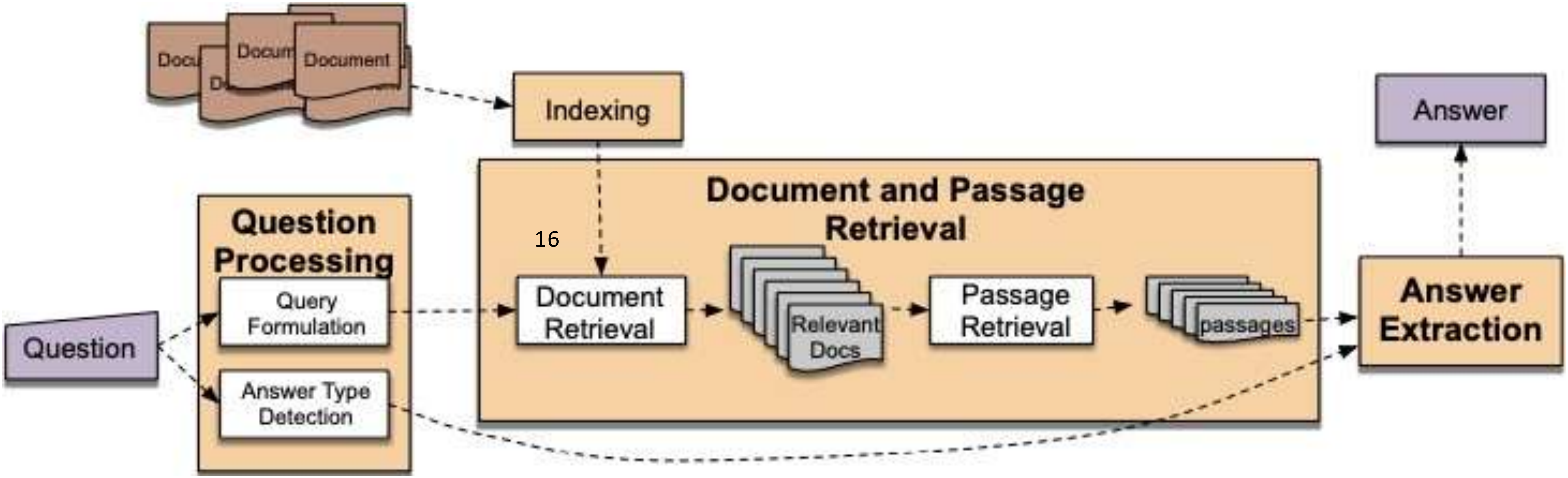
Obtaining labels directly from the environment or simulation



Using human machine iteration to make labelling easier

ARCHITECTURE

Question & Answering System



DATA GATHERING

शिक्षा मंत्रालय
MINISTRY OF
EDUCATION
सत्यमेव जयते

वाणिज्य एवं
उद्योग मंत्रालय
MINISTRY OF
COMMERCE AND INDUSTRY
सत्यमेव जयते

विदेश मंत्रालय
MINISTRY OF
EXTERNAL AFFAIRS
सत्यमेव जयते

सत्यमेव जयते
MOFPI
Ministry of Food Processing Industries
Government of India

PARLIAMENT OF INDIA

पेयजल और स्वच्छता मंत्रालय
MINISTRY OF
DRINKING WATER AND SANITATION
सत्यमेव जयते

कृषि एवं किसान
कल्याण मंत्रालय
MINISTRY OF
AGRICULTURE AND FARMERS WELFARE
सत्यमेव जयते

स्वास्थ्य एवं
परिवार कल्याण मंत्रालय
MINISTRY OF
HEALTH AND FAMILY WELFARE
सत्यमेव जयते

MANY MORE

METHODOLOGY ADOPTED

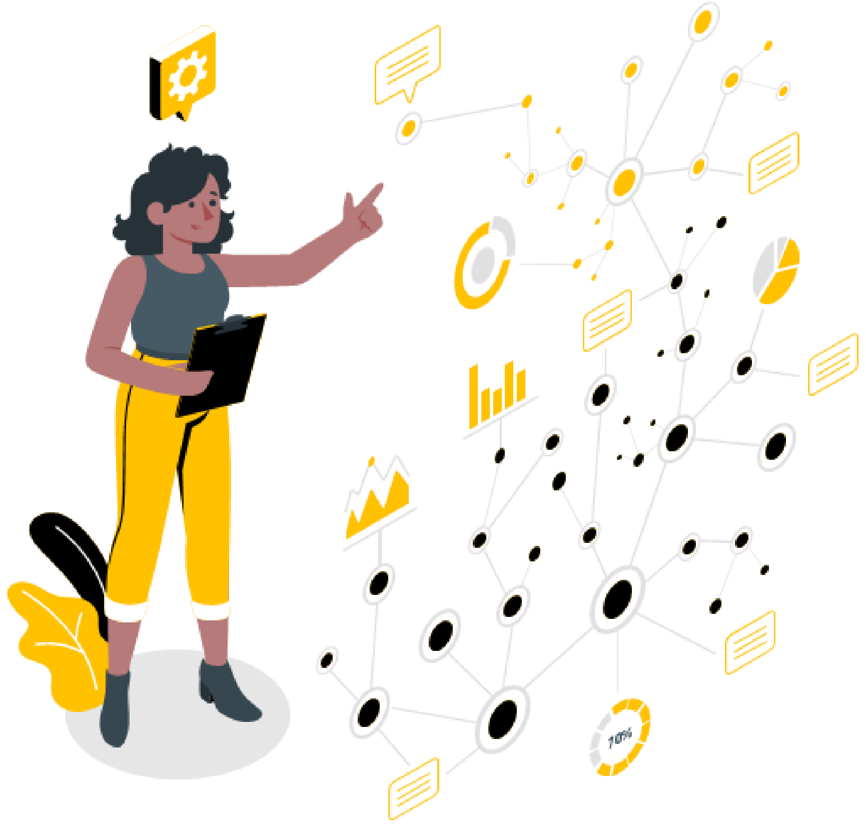
- ✓ Initially, data collection and aggregation was the done for making a QA AI model

- ✓ Required data was collected in the form of titles and paragraphs (maximum 7-15 lines) to provide context for the answers.

- ✓ The data collected was converted to JSON files using user-defined Python functions as a part of the data pre-processing

- ✓ The aim was to create the largest possible amount of data in the restricted time limit.

- ✓ The json file contained the question-answers data in the SQuAD 2.0 format.



END OF PHASE 1

WORK DONE

Generated a total of 8000+ questions from various ministries covererd.

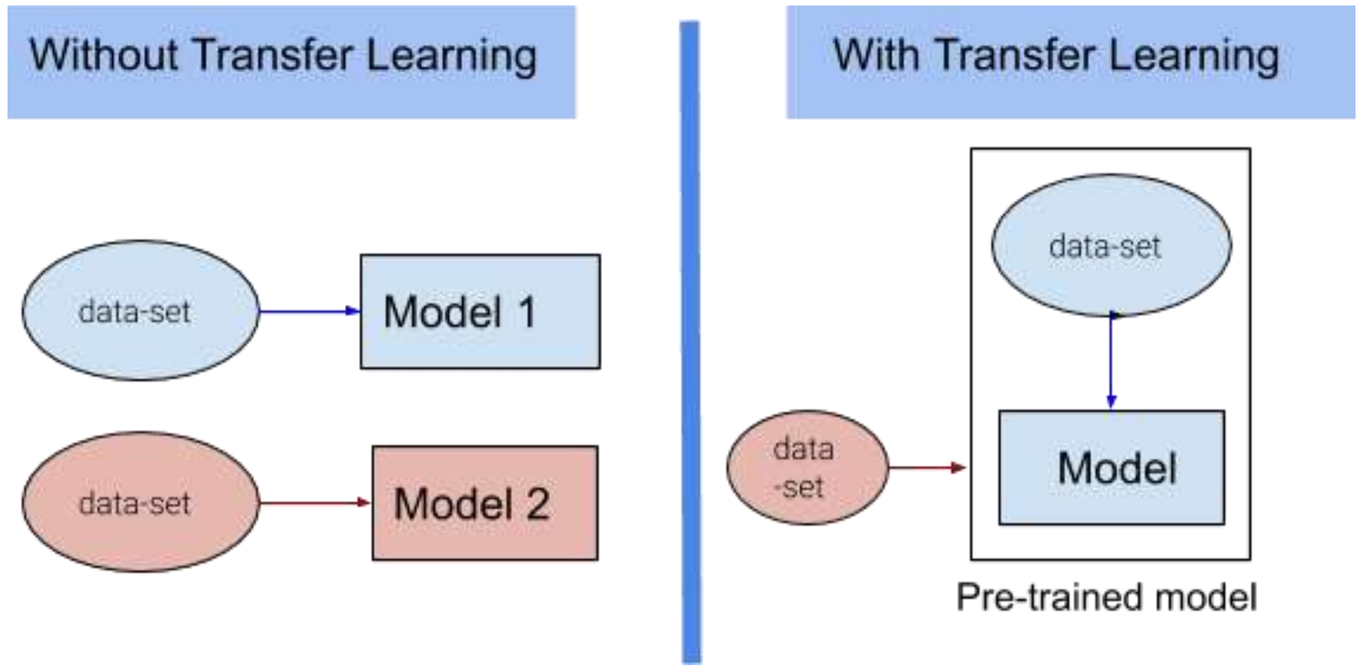
CHALLENGES FACED

- Existing Q&A tools could not generate the level of questions a human can.
- Thus, the script generated for the annotation could not perform up to the mark and had the majority of its questions starting with "What"
- Most of the questions did not make sense.

n for

PHASE II

MODELS & CODE



Transfer Learning

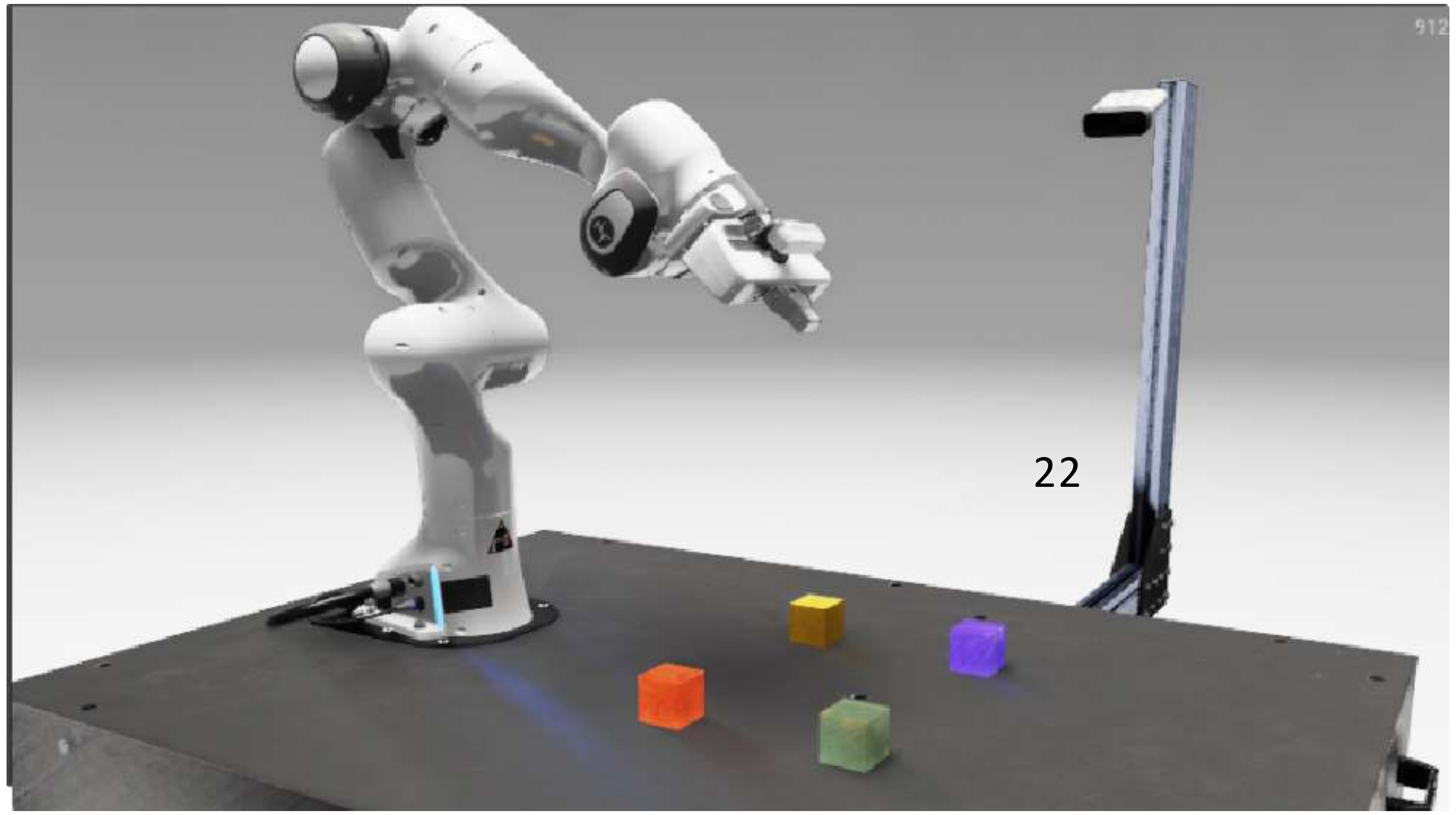
To build any model from scratch, we require a lot of storage and computing power, which may not always be available to us

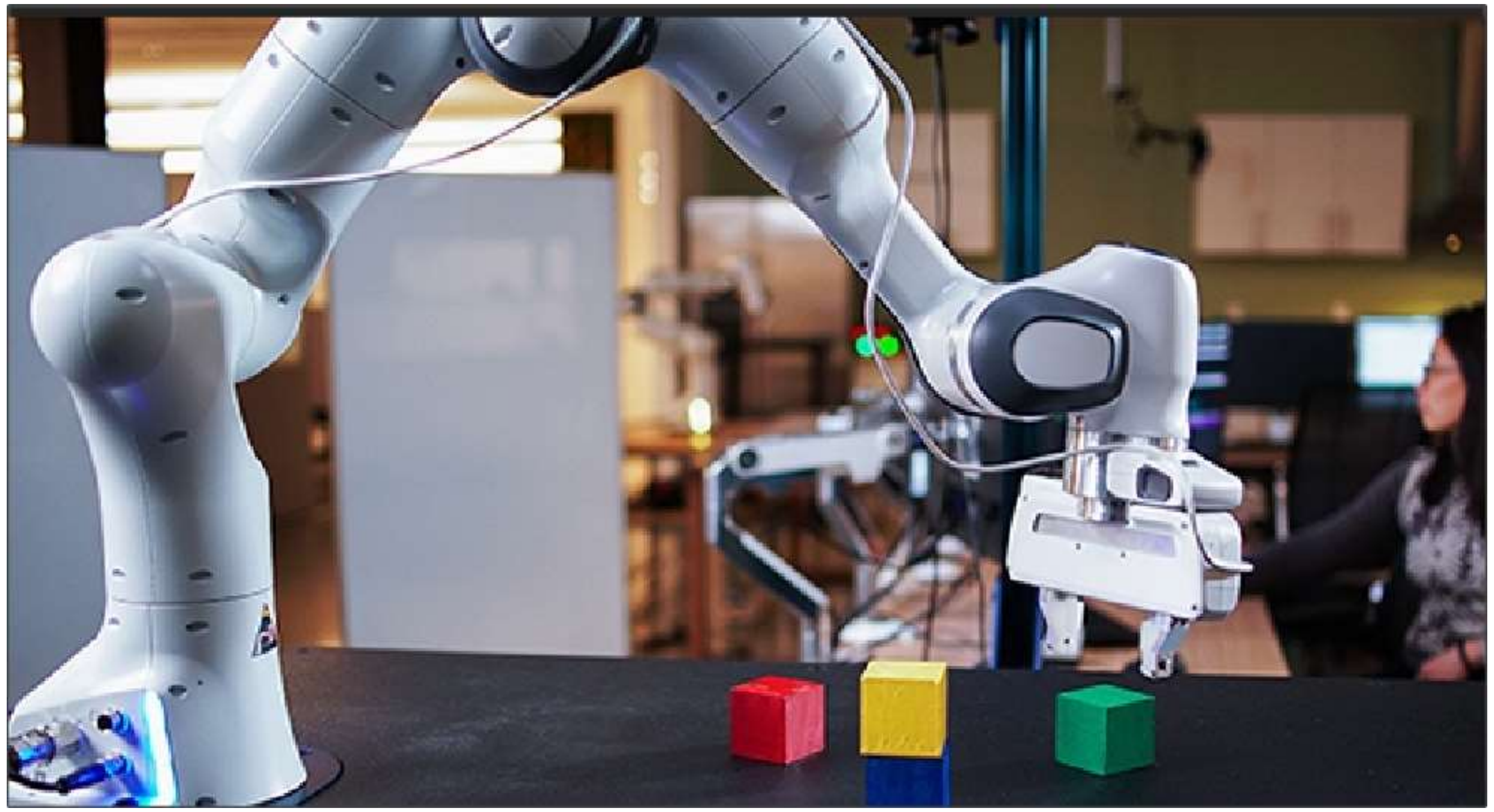
We can face situations where we have methods to improve existing models, but the complications of training the models from scratch once again prevent us from doing so

Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task

However, we must ensure that our task is very similar to the task the pretrained model is meant to do. Otherwise, the result of the tests will not be accurate enough.

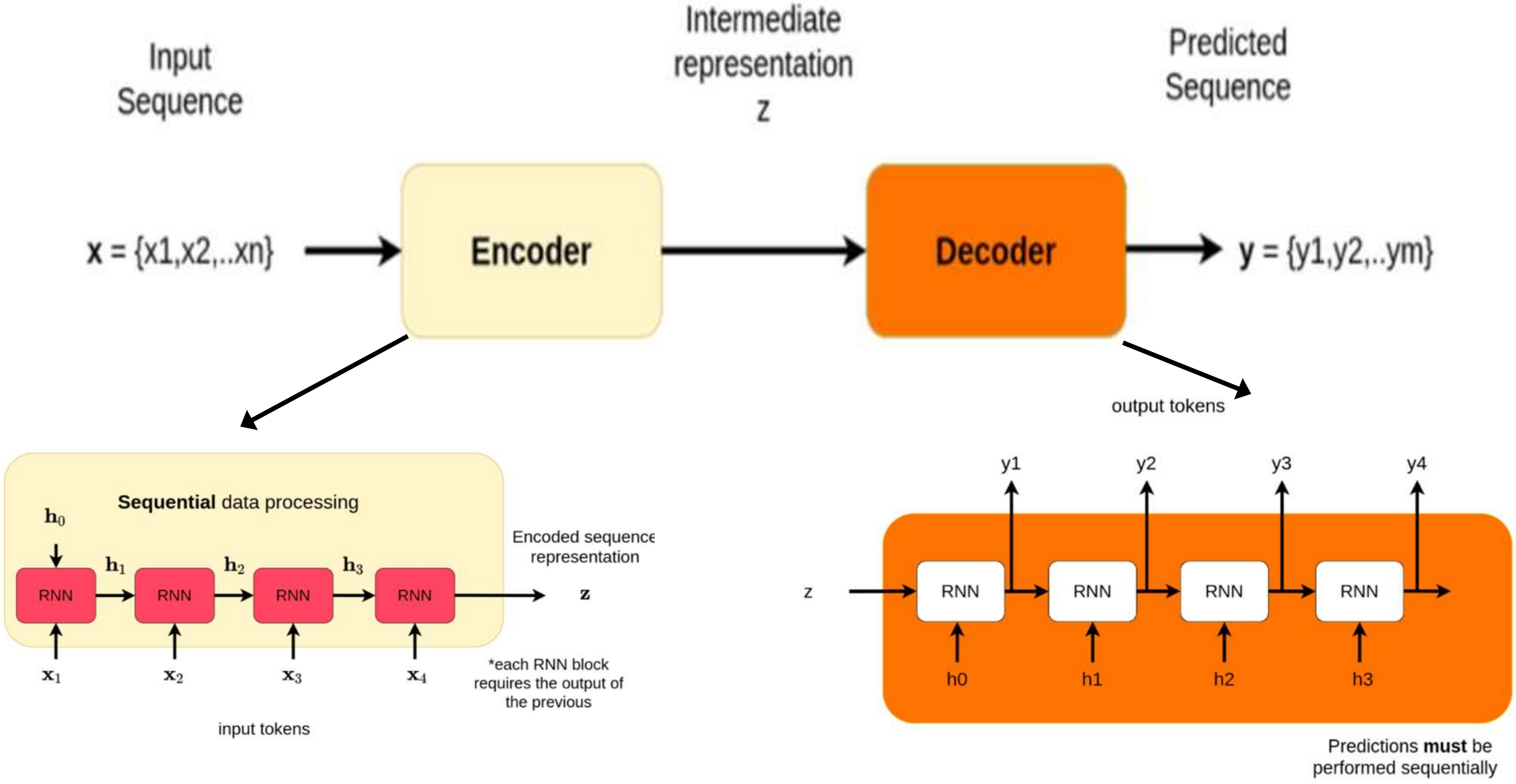
TRANSFER LEARNING: DON’T START FROM SCRATCH





WHICH MODELS TO USE TO MAKE A Q&A TOOL?

Sequence to sequence learning



Input Sequence

Intermediate representation z

Predicted Sequence

Encoded sequence representation

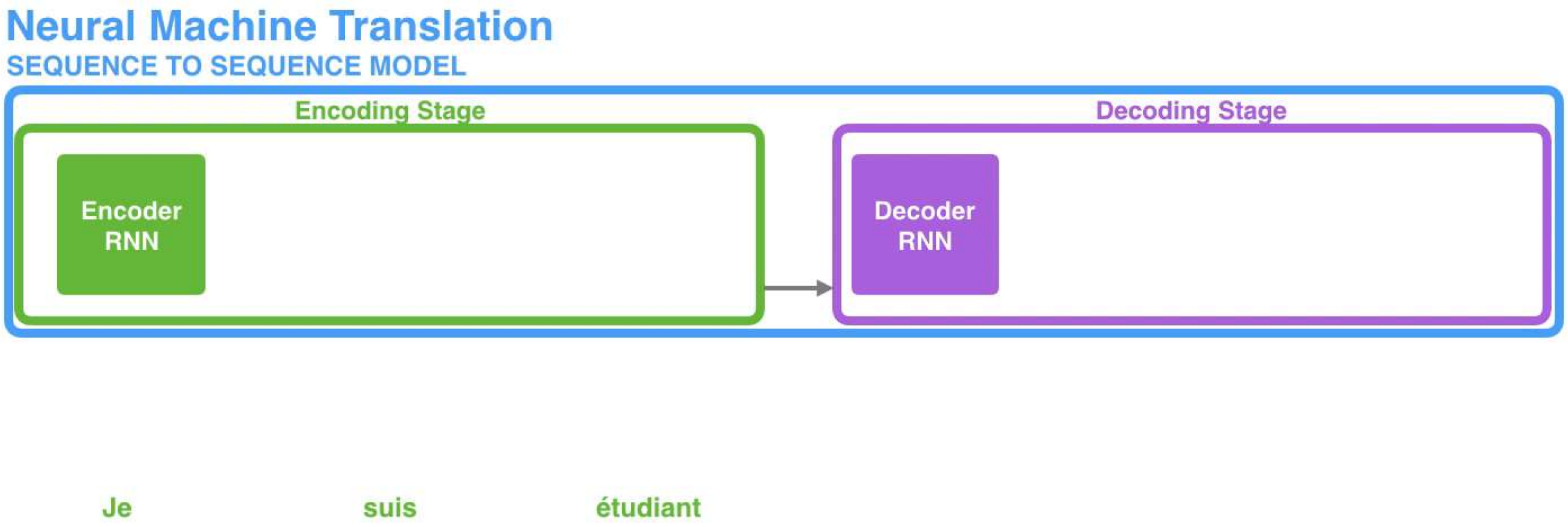
*each RNN block requires the output of the previous

input tokens

output tokens

Predictions **must** be performed sequentially

Sequence to Sequence Working



Neural Machine Translation

SEQUENCE TO SEQUENCE MODEL

Encoding Stage

Encoder RNN

Decoding Stage

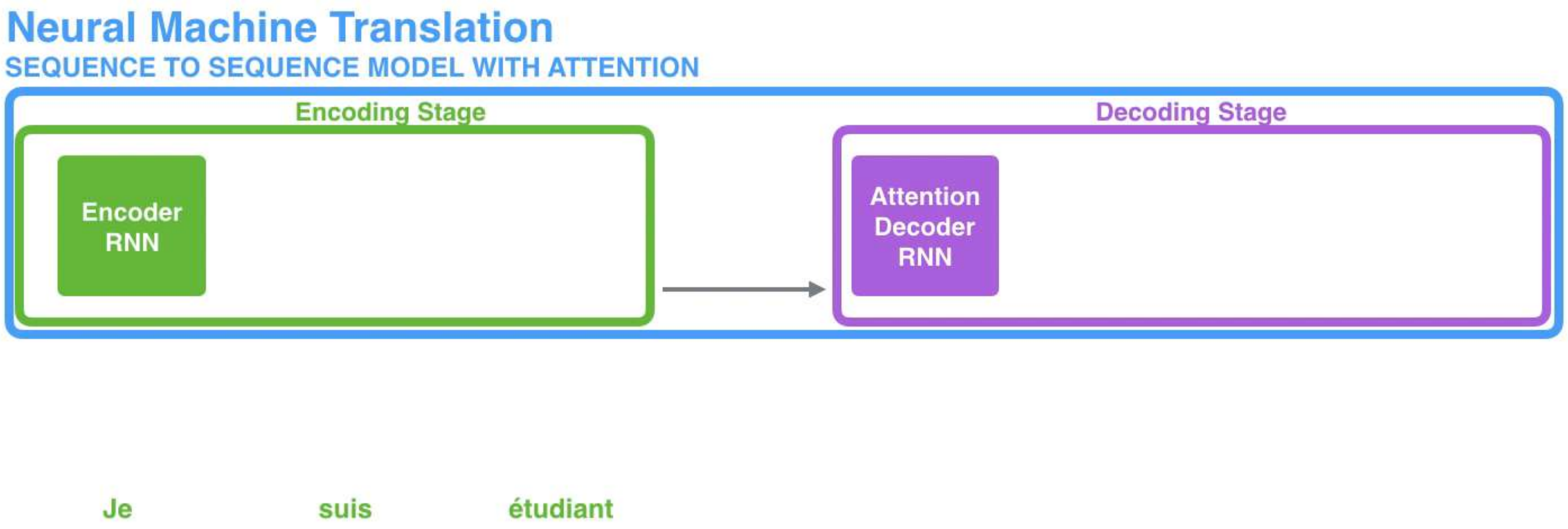
Decoder RNN

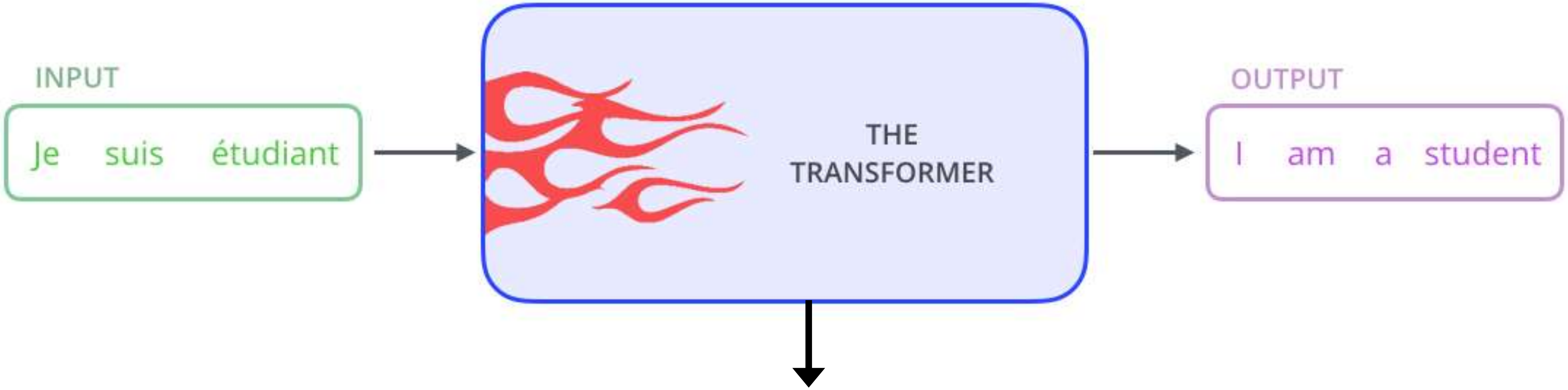
Je

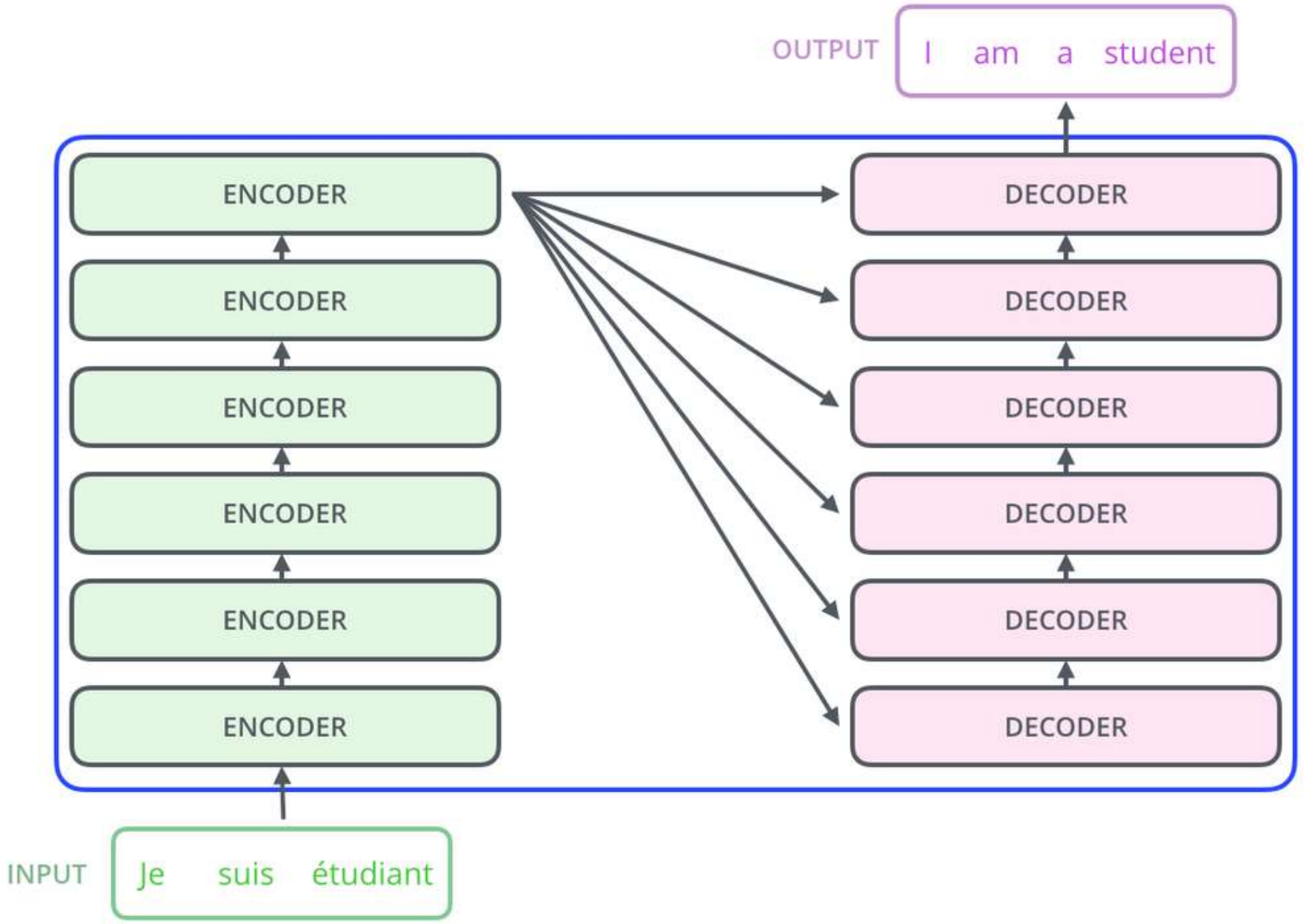
suis

étudiant

Attention is all you need

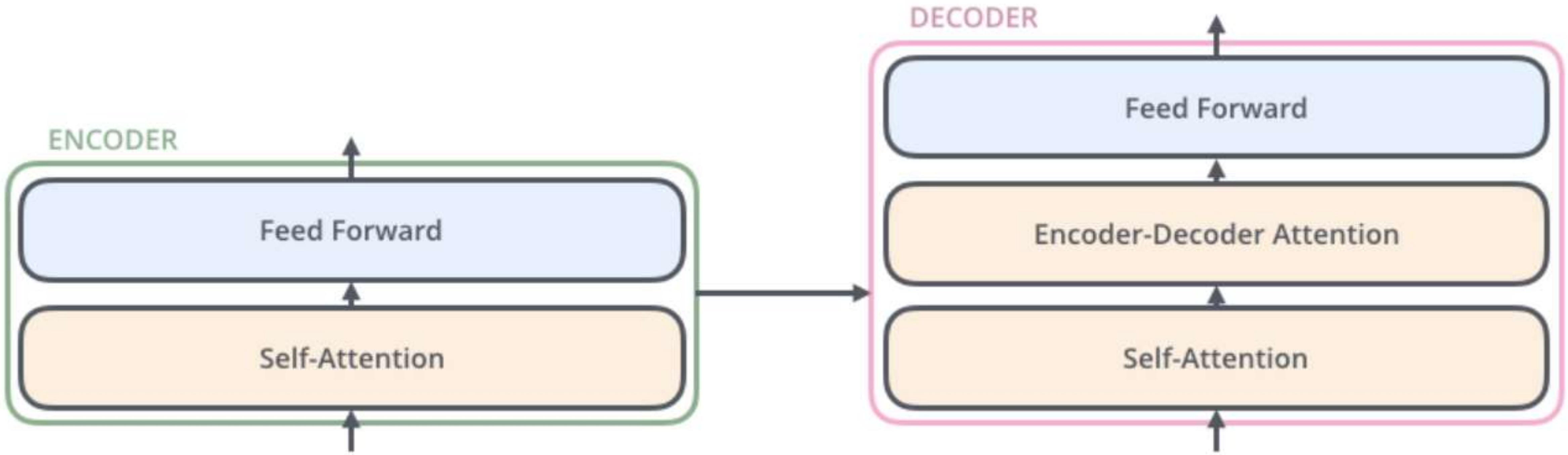






Transformer Architecture

Encoder & Decoder



”THE ANIMAL DIDN'T CROSS THE STREET BECAUSE IT WAS TOO TIRED”

DIFFERENT TRANSFORMER MODELS

BERT

- Bidirectional Encoder Representations

- BERT relies on randomly masking and predicting tokens.

- BERT was specifically trained on Wikipedia (~2.5B words) and Google's BooksCorpus (~800M words)

- Bert was trained on a batch size of 256 sequences

ROBERTA

- Robustly Optimized BERT Pretraining Approach

- dynamically changing the masking pattern applied to the training data

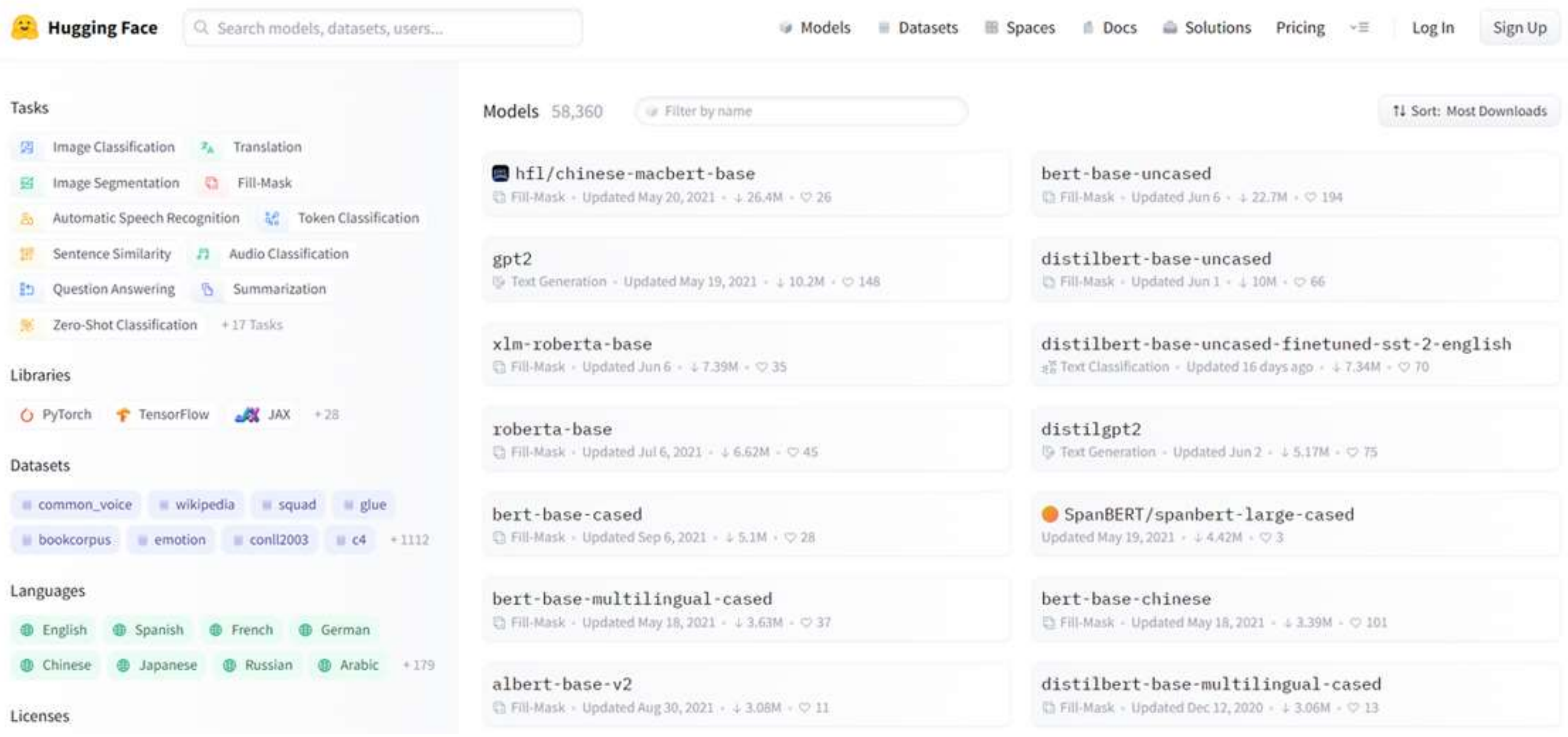
- RoBERTa model was pretrained on the reunion of five datasets which constituted a much larger dataset as compared to BERT

- training on longer sequences and 8k sequences of batch size

HUGGING FACE

About

- Hugging Face is an open-source and platform provider of machine learning technologies. Hugging Face was launched in 2016 and is headquartered in New York City.
- It is a community where all people working in machine learning and AI based technologies help out each other by contributing their models, thereby allowing a user to select a model suitable for his/her work.



Models 58,360

hfl/chinese-macbert-base
Fill-Mask · Updated May 20, 2021 · ↓ 26.4M · ♡ 26

bert-base-uncased
Fill-Mask · Updated Jun 6 · ↓ 22.7M · ♡ 194

gpt2
Text Generation · Updated May 19, 2021 · ↓ 10.2M · ♡ 148

distilbert-base-uncased
Fill-Mask · Updated Jun 1 · ↓ 10M · ♡ 66

xlm-roberta-base
Fill-Mask · Updated Jun 6 · ↓ 7.39M · ♡ 35

distilbert-base-uncased-finetuned-sst-2-english
Text Classification · Updated 16 days ago · ↓ 7.34M · ♡ 70

roberta-base
Fill-Mask · Updated Jul 6, 2021 · ↓ 6.62M · ♡ 45

distilgpt2
Text Generation · Updated Jun 2 · ↓ 5.17M · ♡ 75

bert-base-cased
Fill-Mask · Updated Sep 6, 2021 · ↓ 5.1M · ♡ 28

SpanBERT/spanbert-large-cased
Updated May 19, 2021 · ↓ 4.42M · ♡ 3

bert-base-multilingual-cased
Fill-Mask · Updated May 18, 2021 · ↓ 3.63M · ♡ 37

bert-base-chinese
Fill-Mask · Updated May 18, 2021 · ↓ 3.39M · ♡ 101

albert-base-v2
Fill-Mask · Updated Aug 30, 2021 · ↓ 3.08M · ♡ 11

distilbert-base-multilingual-cased
Fill-Mask · Updated Dec 12, 2020 · ↓ 3.06M · ♡ 13

HYPERPARAMETERS

MAX SEQ LENGTH

The maximum target length to use when predicting with the generate method.

LEARNING RATE

determines the step size at each iteration while moving toward a minimum of a loss function

SAVE STEPS

Determines the checkpointing of model after fixed specified steps

BATCH SIZE

Refers to the number of training examples utilized in one iteration

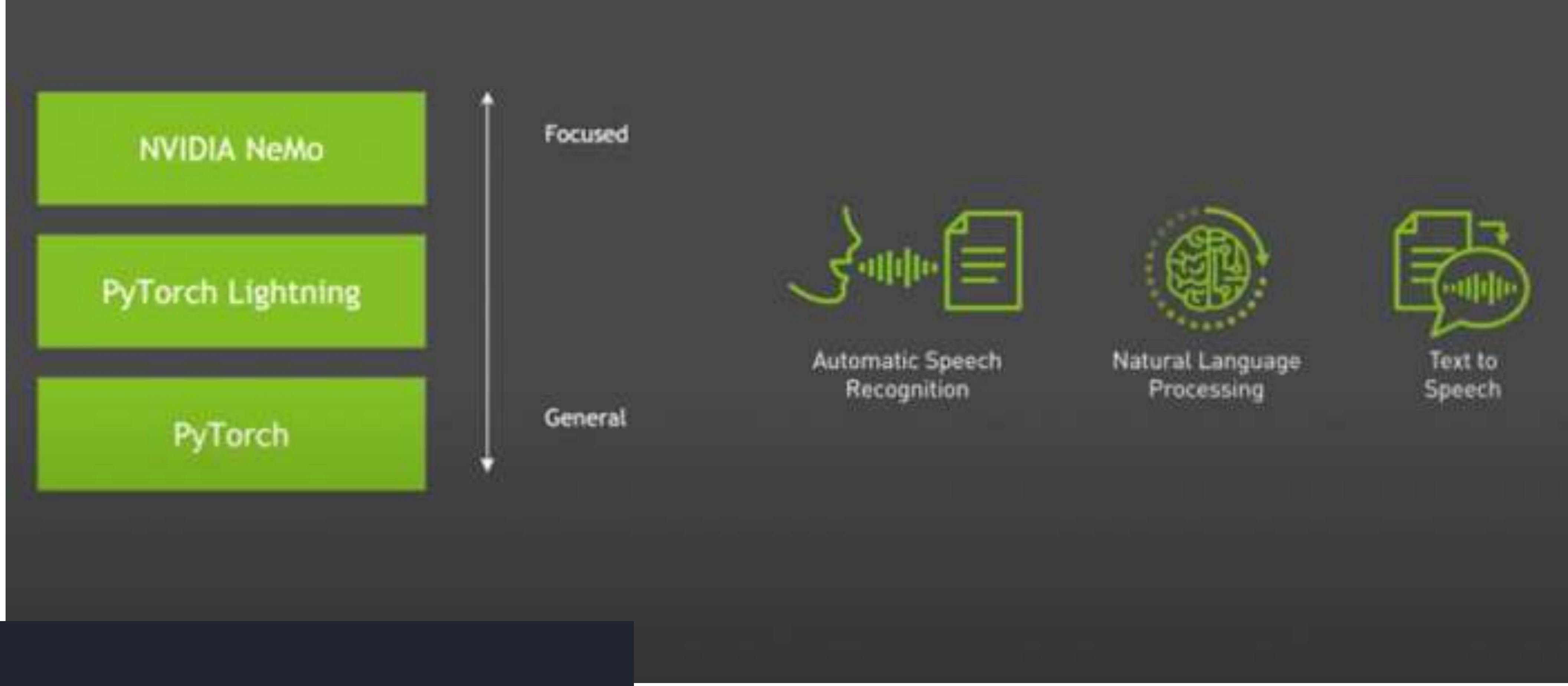
EPOCH

indicates the number of passes of the entire training dataset the machine learning algorithm has completed

DOC_STRIDE

Modifies the amount of movement over the tokenized text

NVDIA NEMO



- NVIDIA NeMo (Neural Modules), part of the NVIDIA AI platform, is a toolkit for building new state-of-the-art conversational AI models.

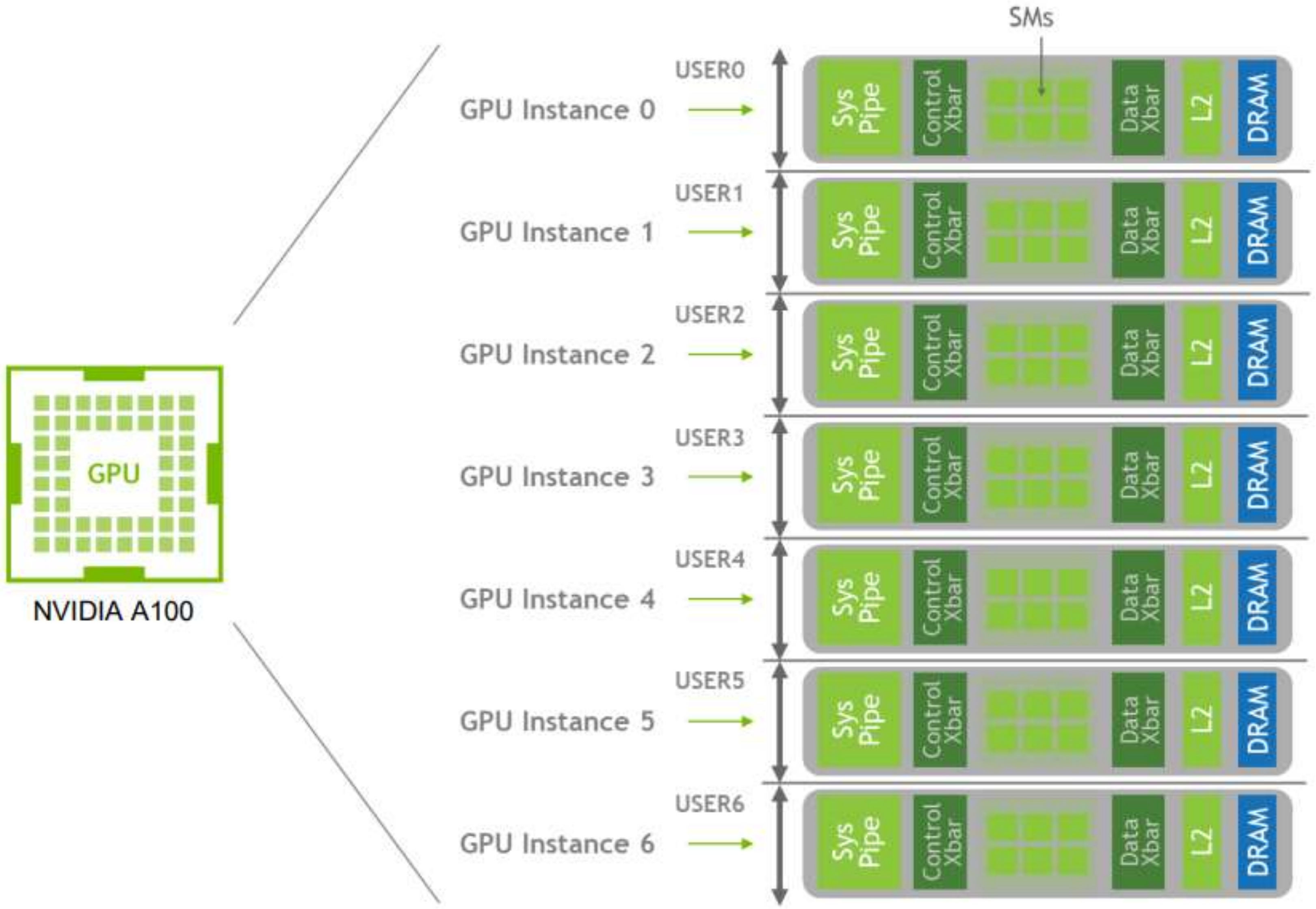
- NeMo has separate collections for Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and Text-to-Speech (TTS) models.

- Each collection consists of prebuilt modules that include everything needed to train on your data. Every module can easily be customized, extended, and composed to create new conversational AI model architectures.

- The NIC AI Hackathon 2022 was powered by NVIDIA and training and testing was done using their most powerful cloud GPU platform.

NVIDIA.
NEMO

MULTI-INSTANCE GPU (MIG)



Profile Name	# Instances per GPU	Fraction of Memory	Fraction of Compute (SMs)	Hardware Units	Target Workload (Use-cases are inclusive)
MIG 1g.5gb	7	1/8	1/7	0 NVDECs	Jupyter Notebooks for Development, Matlab, Model Tuning, Inference, Light HPC
MIG 2g.10gb	3	2/8	2/7	1 NVDEC	Inference, Light HPC
MIG 3g.20gb	2	4/8	3/7	2 NVDECs	Light Training, Inference, Light HPC
MIG 4g.20gb	1	4/8	4/7	2 NVDECs	Light Training, Inference, Light HPC
MIG 7g.40gb	1	Full	7/7	5 NVDECs / OFA / NVJPG	Training, Light HPC

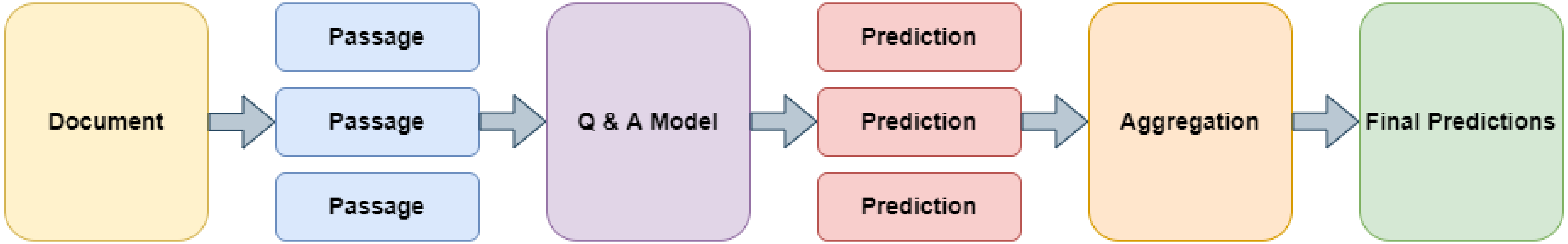
METHODOLOGY USED FOR TRAINING AI MODEL

- ✓ The data from all the ministries were taken and split into train, test , and validation(if needed)

- ✓ Various experimentations were done by changing hyperparameters and models to find the most optimal results i.e. until signs of overfitting were discovered, i.e. till the accuracy graph of our model maintained a gradual ascent.

- ✓ The training was extensive and hence each epoch was done on a NVIDIA remote GPU computing platform to make the huge NLP task faster.

The Project Architecture



EVALUATION

01 Exact Match (EM):

For each question-answer pair, if the characters of the model's prediction exactly match the characters of (one of) the True Answer(s), EM = 1, otherwise EM = 0

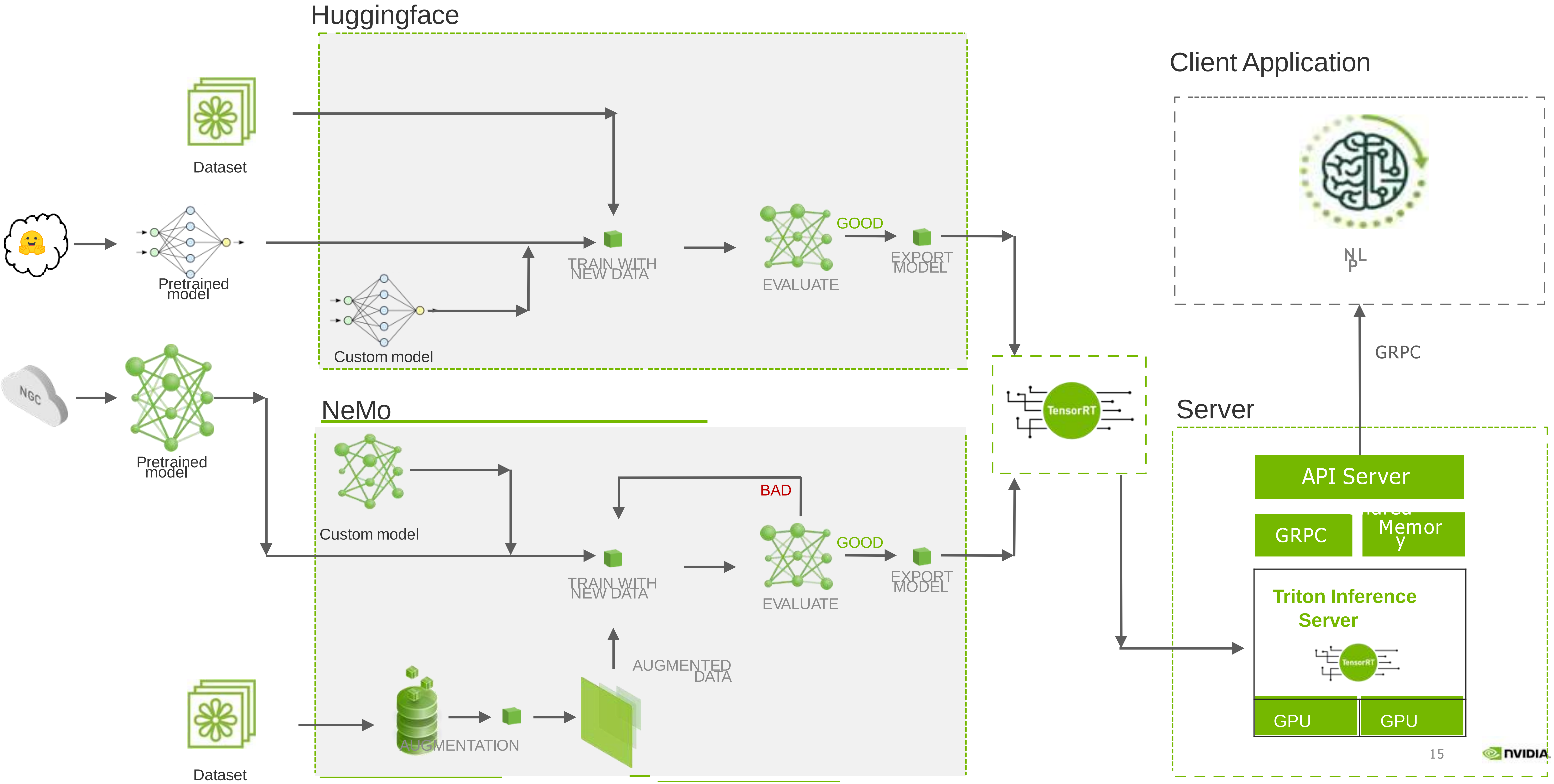
02 F-1 Score (Macro-averaged

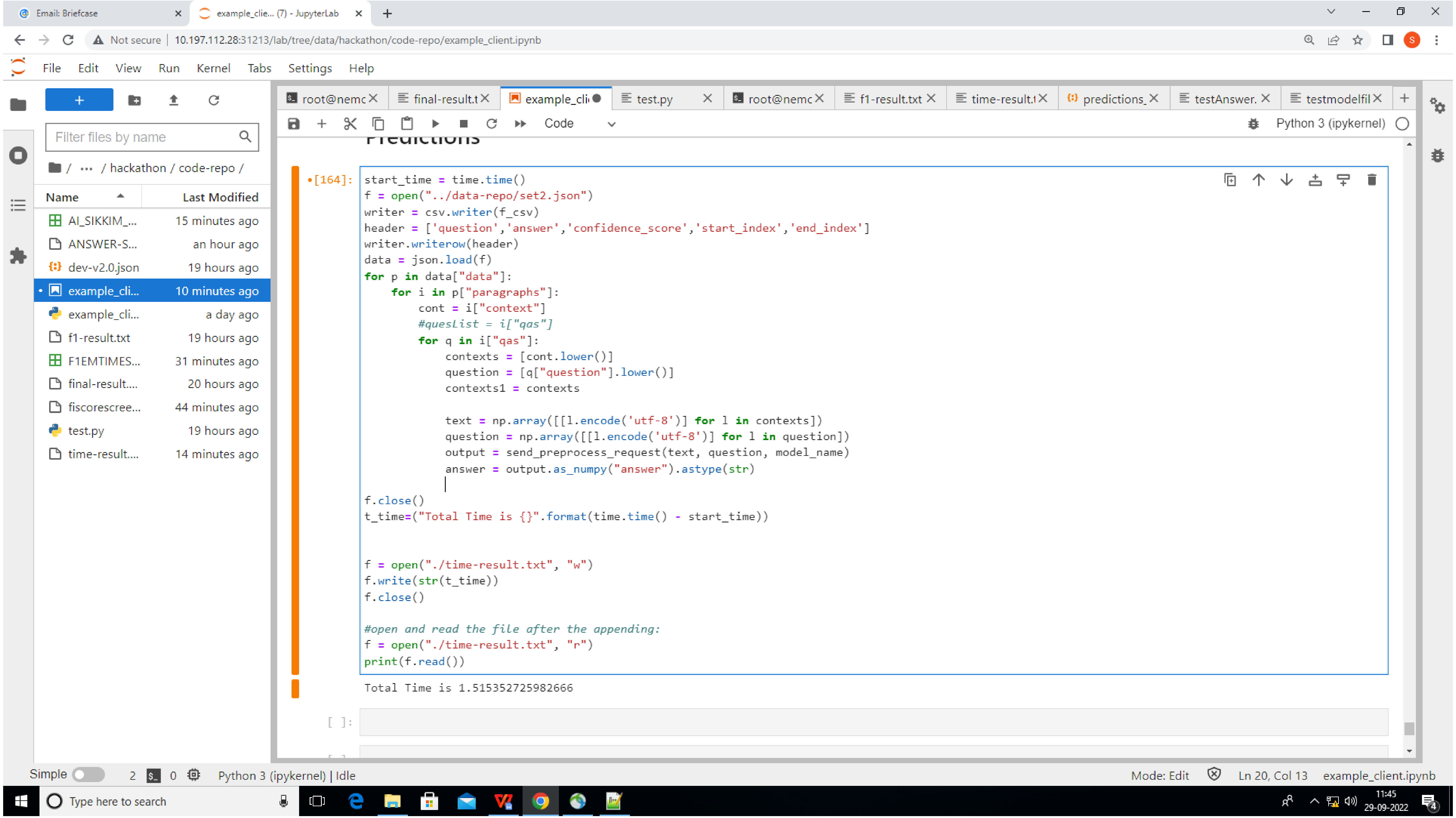
$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

The number of shared words between the prediction and the truth is the basis of the F1 score: precision is the ratio of the number of shared words to the total number of words in the prediction, and recall is the ratio of the number of shared words to the total number of words in the ground truth.

NLP AI PLATFORM

Huggingface(training)+ NeMo (training) + Triton(inference)





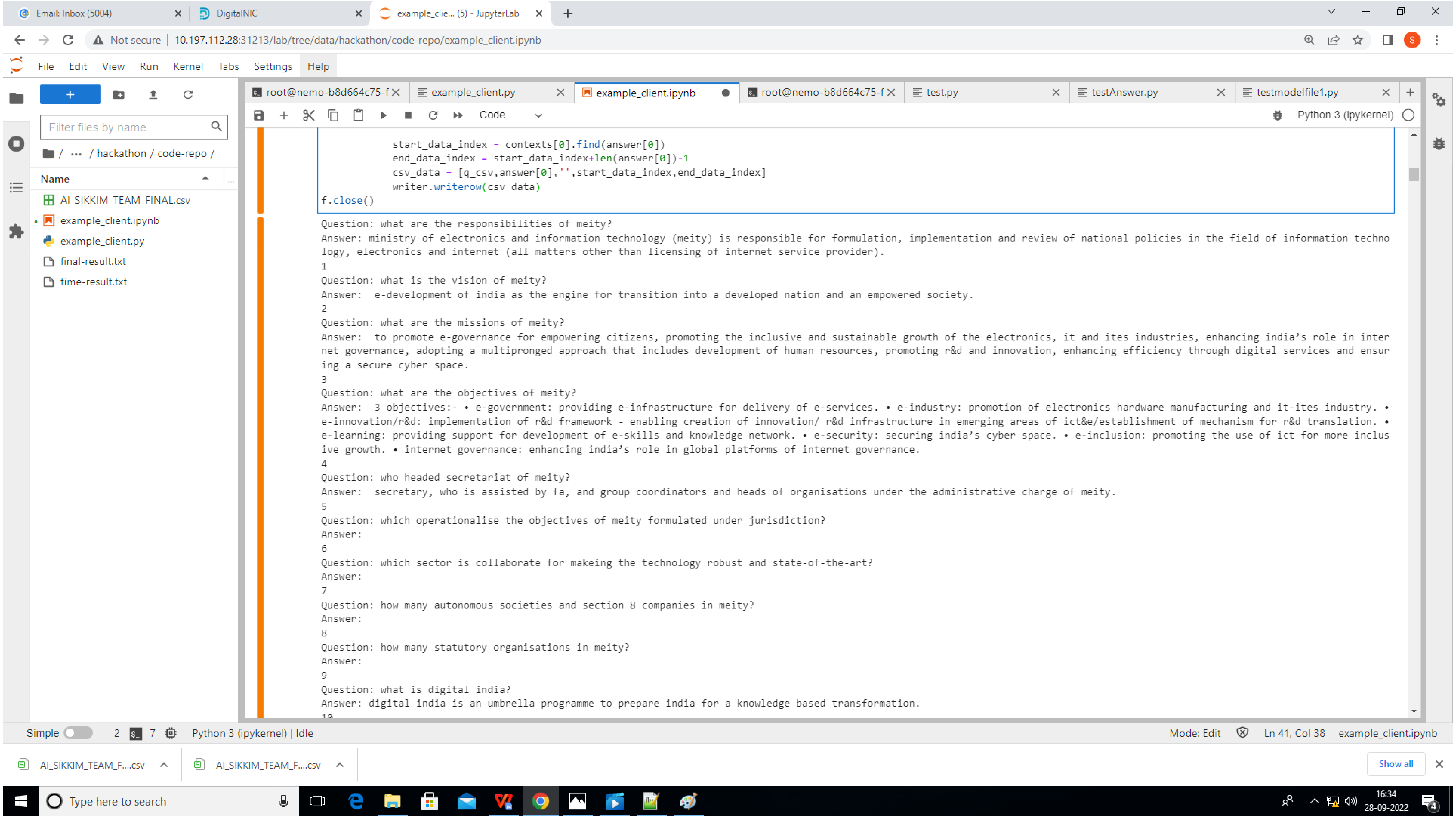
Name	Last Modified
AI_SIKKIM_...	15 minutes ago
ANSWER-S...	an hour ago
dev-v2.0.json	19 hours ago
example_cli...	10 minutes ago
example_cli...	a day ago
f1-result.txt	19 hours ago
F1EMTIMES...	31 minutes ago
final-result....	20 hours ago
fiscorescree...	44 minutes ago
test.py	19 hours ago
time-result....	14 minutes ago

•[164]:

```
start_time = time.time()
f = open("../data-repo/set2.json")
writer = csv.writer(f_csv)
header = ['question','answer','confidence_score','start_index','end_index']
writer.writerow(header)
data = json.load(f)
for p in data["data"]:
    for i in p["paragraphs"]:
        cont = i["context"]
        #quesList = i["qas"]
        for q in i["qas"]:
            contexts = [cont.lower()]
            question = [q["question"].lower()]
            contexts1 = contexts

            text = np.array([[l.encode('utf-8')] for l in contexts])
            question = np.array([[l.encode('utf-8')] for l in question])
            output = send_preprocess_request(text, question, model_name)
            answer = output.as_numpy("answer").astype(str)

f.close()
t_time=("Total Time is {}".format(time.time() - start_time))


f = open("./time-result.txt", "w")
f.write(str(t_time))
f.close()

#open and read the file after the appending:
f = open("./time-result.txt", "r")
print(f.read())
```

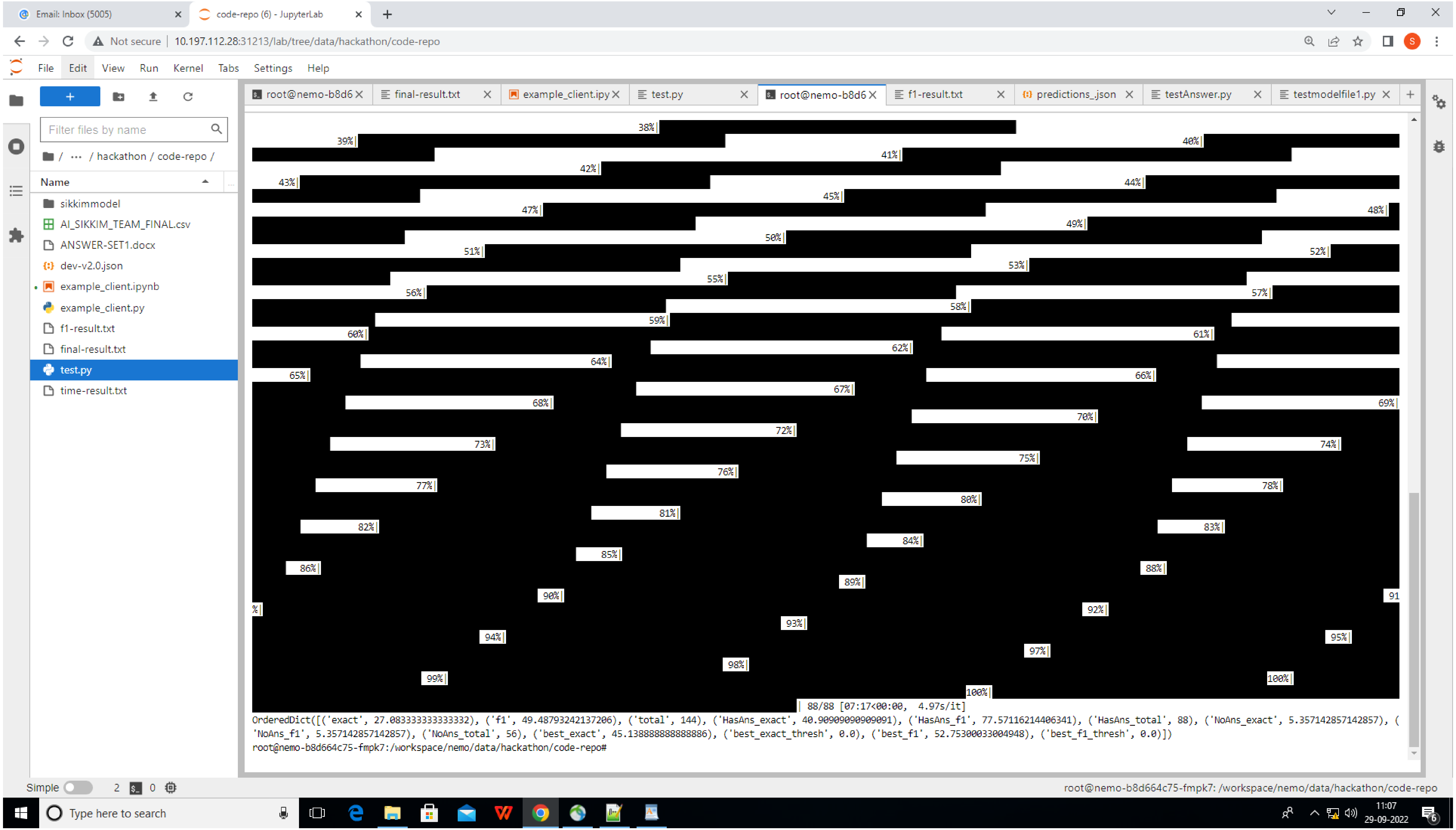
```
Total Time is 1.515352725982666
```

[ ]:



```
            start_data_index = contexts[0].find(answer[0])
            end_data_index = start_data_index+len(answer[0])-1
            csv_data = [q_csv,answer[0],'',start_data_index,end_data_index]
            writer.writerow(csv_data)
f.close()
```

```
Question: what are the responsibilities of meity?
Answer: ministry of electronics and information technology (meity) is responsible for formulation, implementation and review of national policies in the field of information techno
logy, electronics and internet (all matters other than licensing of internet service provider).
1
Question: what is the vision of meity?
Answer:  e-development of india as the engine for transition into a developed nation and an empowered society.
2
Question: what are the missions of meity?
Answer:  to promote e-governance for empowering citizens, promoting the inclusive and sustainable growth of the electronics, it and ites industries, enhancing india’s role in inter
net governance, adopting a multipronged approach that includes development of human resources, promoting r&d and innovation, enhancing efficiency through digital services and ensur
ing a secure cyber space.
3
Question: what are the objectives of meity?
Answer:  3 objectives:- • e-government: providing e-infrastructure for delivery of e-services. • e-industry: promotion of electronics hardware manufacturing and it-ites industry. •
e-innovation/r&d: implementation of r&d framework - enabling creation of innovation/ r&d infrastructure in emerging areas of ict&e/establishment of mechanism for r&d translation. •
e-learning: providing support for development of e-skills and knowledge network. • e-security: securing india’s cyber space. • e-inclusion: promoting the use of ict for more inclus
ive growth. • internet governance: enhancing india’s role in global platforms of internet governance.
4
Question: who headed secretariat of meity?
Answer:  secretary, who is assisted by fa, and group coordinators and heads of organisations under the administrative charge of meity.
5
Question: which operationalise the objectives of meity formulated under jurisdiction?
Answer:
6
Question: which sector is collaborate for makeing the technology robust and state-of-the-art?
Answer:
7
Question: how many autonomous societies and section 8 companies in meity?
Answer:
8
Question: how many statutory organisations in meity?
Answer:
9
Question: what is digital india?
Answer: digital india is an umbrella programme to prepare india for a knowledge based transformation.
10
```



```
38%|
39%|                                    40%|
41%|
42%|
43%|                                    44%|
45%|
47%|                                    48%|
49%|
50%|
51%|                                    52%|
53%|
55%|
56%|                                    57%|
58%|
59%|
60%|                                    61%|
62%|
64%|
65%|                                    66%|
67%|
68%|                                    69%|
70%|
72%|
73%|                                    74%|
75%|
76%|
77%|                                    78%|
80%|
81%|
82%|                                    83%|
84%|
85%|
86%|                                    88%|
89%|
90%|                                    91
%|                                      92%|
93%|
94%|                                    95%|
97%|
98%|
99%|                                    100%|
100%|
| 88/88 [07:17<00:00,  4.97s/it]
OrderedDict([('exact', 27.083333333333332), ('f1', 49.48793242137206), ('total', 144), ('HasAns_exact', 40.90909090909091), ('HasAns_f1', 77.57116214406341), ('HasAns_total', 88), ('NoAns_exact', 5.357142857142857), (
'NoAns_f1', 5.357142857142857), ('NoAns_total', 56), ('best_exact', 45.138888888888886), ('best_exact_thresh', 0.0), ('best_f1', 52.75300033004948), ('best_f1_thresh', 0.0)])
root@nemo-b8d664c75-fmpk7:/workspace/nemo/data/hackathon/code-repo#
```

KEY HIGHLIGHTS & LEARNINGS

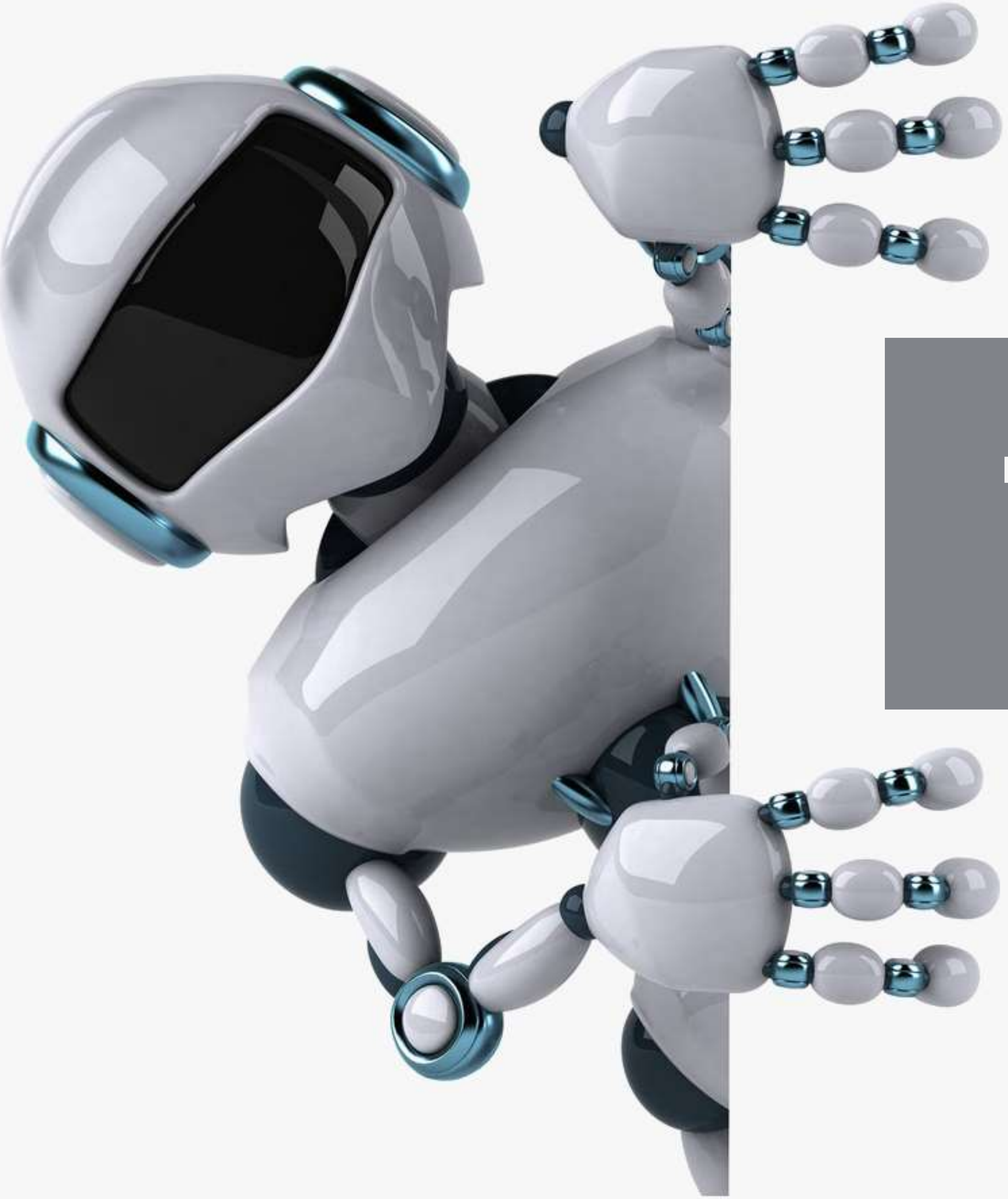
1 Meta data collection and annotation and SQuAD 2.0 format.

2 Architecture of Transformers and the reasons why it is the best option to use.

3 Details about models like BERT, RoBERTa, and XLNET.

4 Cloud platforms like NVIDIA NeMo and their essence while handling such large datasets.

5 Importance of Pretrained models and Transfer Learning.



THANK YOU