LEAP powered by Intel® oneAPI AI Analytics Toolkit

Problem Statement : **Open Innovation in Education**

Team Name : C5ailabs

Team Members : Rohit Sroch, Sujith R Kumar, Mohan K Rachumallu, Shubham Jain

Problem Statement

MOOCs
(Massive Open Online Courses)

200K
Users in 2012

380M
Users in 2020

34.26%
CAGR 2022-27*

5% -10%
Completion rate

Key Challenges

- Lengthy videos
- Instructor Availability
- Slow response from forums
- No real time Q&A/Mentor

Approach

LEAP

(Learning Enhancement and Assistance Platform)

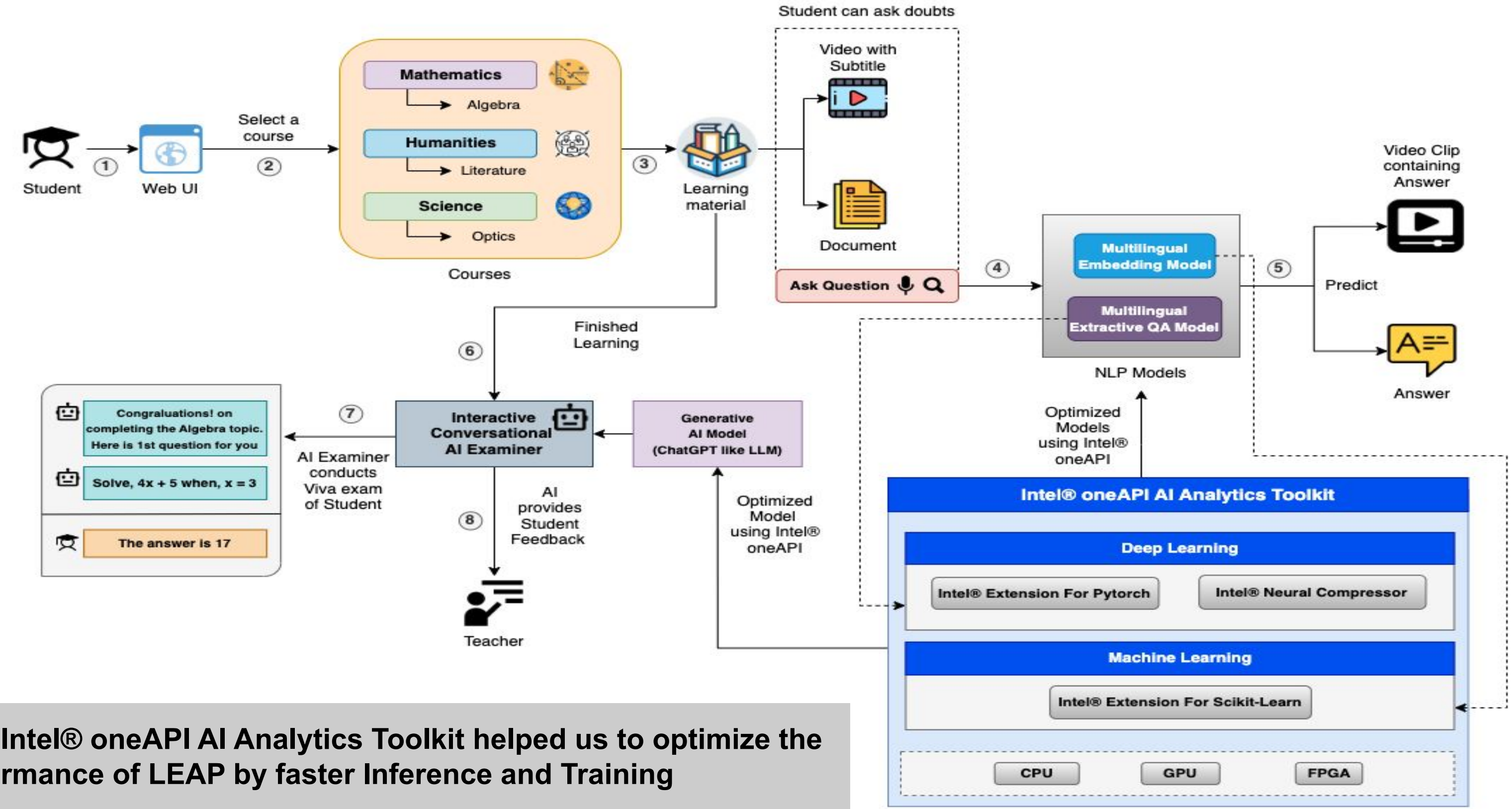
- AI based platform
- Powered by Intel OneAPI
- Quality Education
- All time Availability

Key Features of LEAP

- **Ask Question/Doubt**
- **Conversational AI Examiner**
- **Feedback from AI Examiner**
- **Multilingual Support**

Reference: PRNewswire , Edtechreview; holoniq

High Level Architecture


Student can ask doubts
Video with Subtitle
Document
Ask Question
Mathematics
Algebra
Humanities
Literature
Science
Optics
Courses
Student
Web UI
Select a course
1
2
3
Learning material
4
5
Multilingual Embedding Model
Multilingual Extractive QA Model
NLP Models
Predict
Video Clip containing Answer
Answer
Finished Learning
6
7
Interactive Conversational AI Examiner
Generative AI Model (ChatGPT like LLM)
AI Examiner conducts Viva exam of Student
8
AI provides Student Feedback
Teacher
Optimized Model using Intel® oneAPI
Optimized Models using Intel® oneAPI
Intel® oneAPI AI Analytics Toolkit
Deep Learning
Intel® Extension For Pytorch
Intel® Neural Compressor
Machine Learning
Intel® Extension For Scikit-Learn
CPU
GPU
FPGA

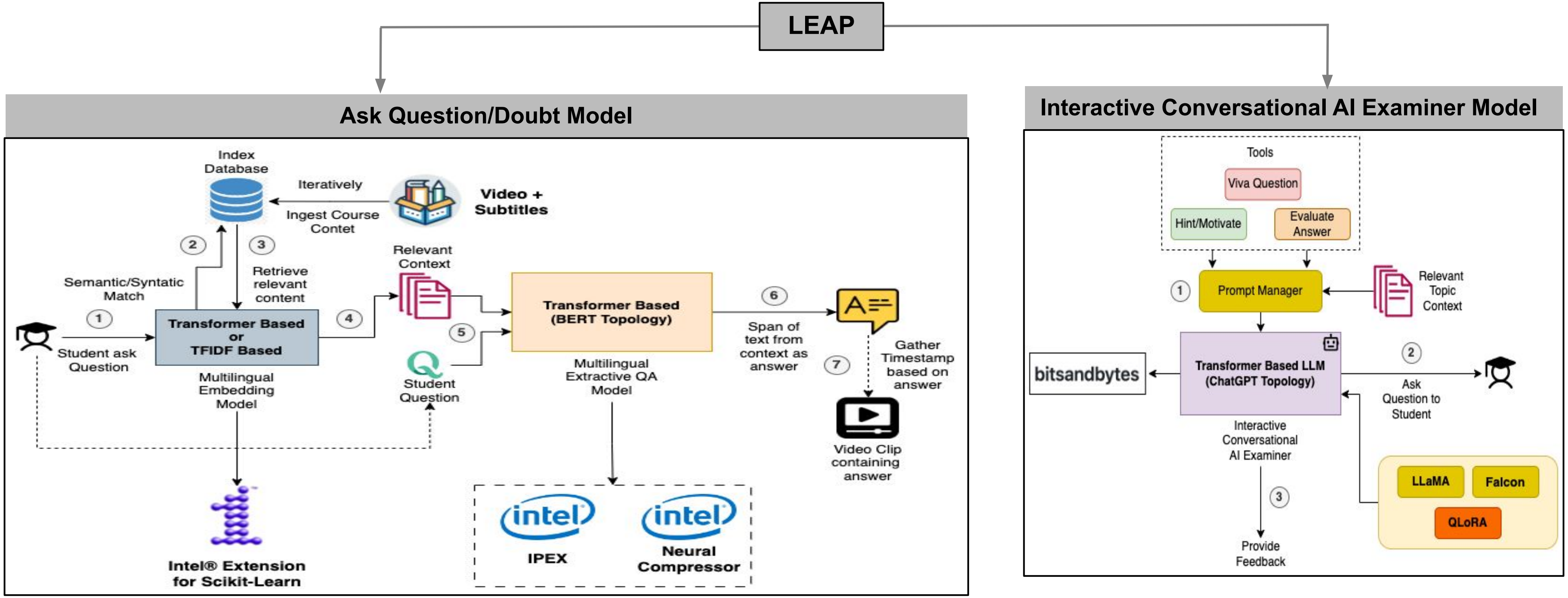
Congraluations! on completing the Algebra topic. Here is 1st question for you

Solve, 4x + 5 when, x = 3

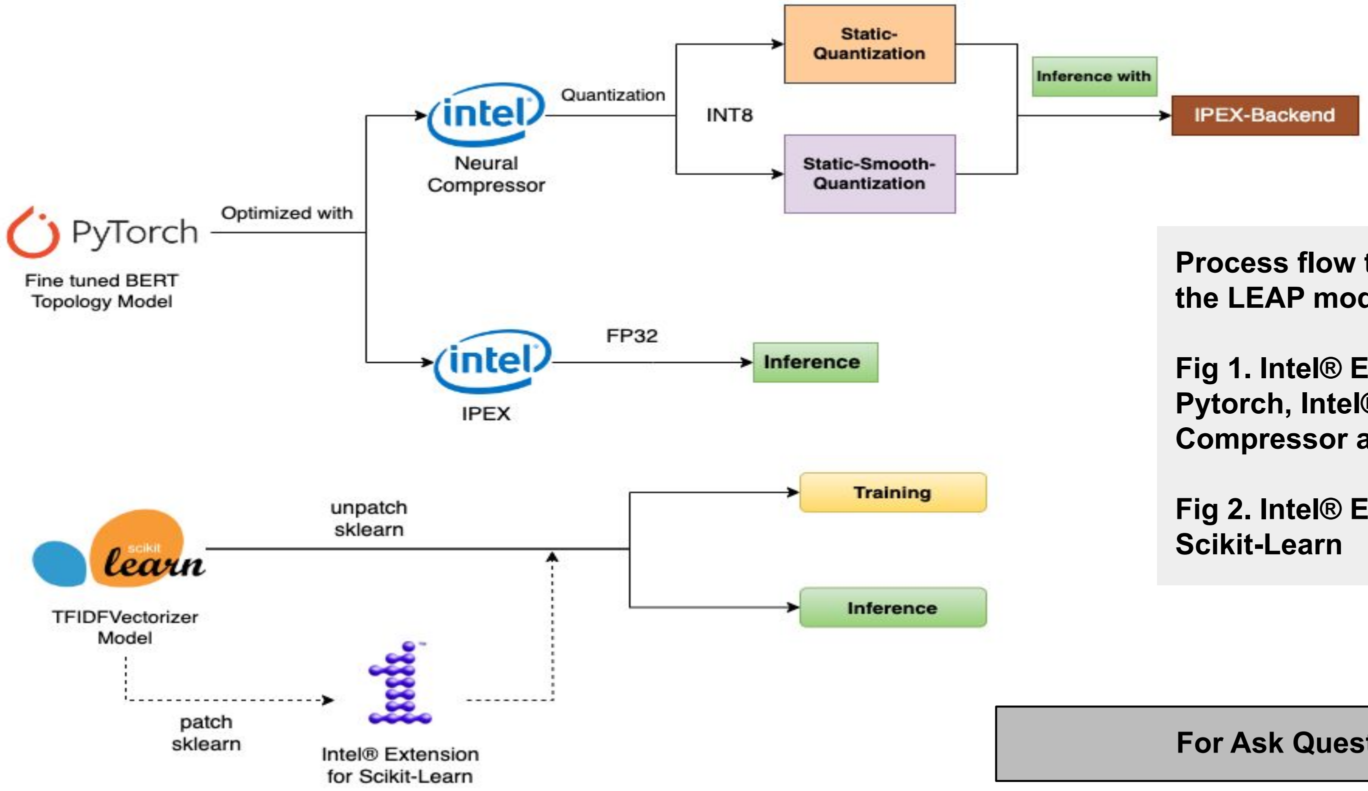
The answer is 17

Here Intel® oneAPI AI Analytics Toolkit helped us to optimize the performance of LEAP by faster Inference and Training

LEAP: Detailed Model Architecture Diagram for Both Components


LEAP
Ask Question/Doubt Model
Index Database
Iteratively
Ingest Course Contet
Video + Subtitles
2
3
Semantic/Syntatic Match
Retrieve relevant content
Relevant Context
1
Student ask Question
Transformer Based or TFIDF Based
Multilingual Embedding Model
4
5
Student Question
Transformer Based (BERT Topology)
Multilingual Extractive QA Model
6
Span of text from context as answer
7
Gather Timestamp based on answer
Video Clip containing answer
Intel® Extension for Scikit-Learn
intel
IPEX
intel
Neural Compressor
Interactive Conversational AI Examiner Model
Tools
Viva Question
Hint/Motivate
Evaluate Answer
1
Prompt Manager
Relevant Topic Context
bitsandbytes
Transformer Based LLM (ChatGPT Topology)
2
Ask Question to Student
Interactive Conversational AI Examiner
3
Provide Feedback
LLaMA
Falcon
QLoRA

Result Summary (unique aspects of oneAPI/SYCL used)



Process flow to optimize the LEAP models by using

Fig 1. Intel® Extension for Pytorch, Intel® Neural Compressor and

Fig 2. Intel® Extension for Scikit-Learn

For Ask Question/Doubt

Static-QAT-INT8 is 4.44x times faster in performance as compared to baseline For our Extractive QA Model

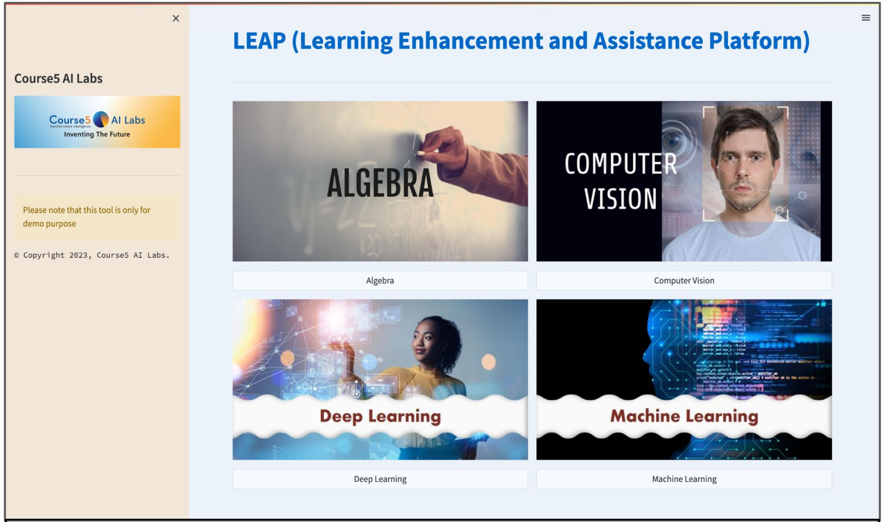
For Ask Question/Doubt				
Extractive Question Answering Model				
	Pytorch (Base) - FP32	Pytorch (IPEX) - FP32	**Static-QAT-INT8**	Static-Smooth-QAT-INT8
Latency (milli sec)	64.513	39.329	**14.514**	15.24
Throughput (samples/sec)	15.501	25.427	**68.9**	65.616
F1 Score (SQuAD-v1)	**76.11**	**76.11**	75.72	75.72

For Ask Question/Doubt		
TFIDF Embedding Model		
	Scikit-Learn (Base)	**Intel Extension For Scikit-Learn**
Latency (milli sec)	0.761	**0.752**
Throughput (samples/sec)	1313.63	**1330.49**

*Table: Latency/Throughput/Speed-Up Benchmark result for **our Extractive Question Answering ALBERT Model (Multilingual) and TFIDF Embedding Model** on Intel® Dev Cloud machine (**Intel® Xeon® Platinum 8480+ (4th Gen: Sapphire Rapids) - 224v CPUs 503GB RAM**) with optimization using IPEX-FP32, Static-QAT-INT8 using Intel® Neural Compressor and TFIDFVectorizer using Intel® Extension for Scikit-Learn.*

Demo Link and Screenshots

Link: https://www.youtube.com/watch?v=M51BFcoJa3k



Course5 AI Labs

Please note that this tool is only for demo purpose

© Copyright 2023, Course5 AI Labs.

LEAP (Learning Enhancement and Assistance Platform)

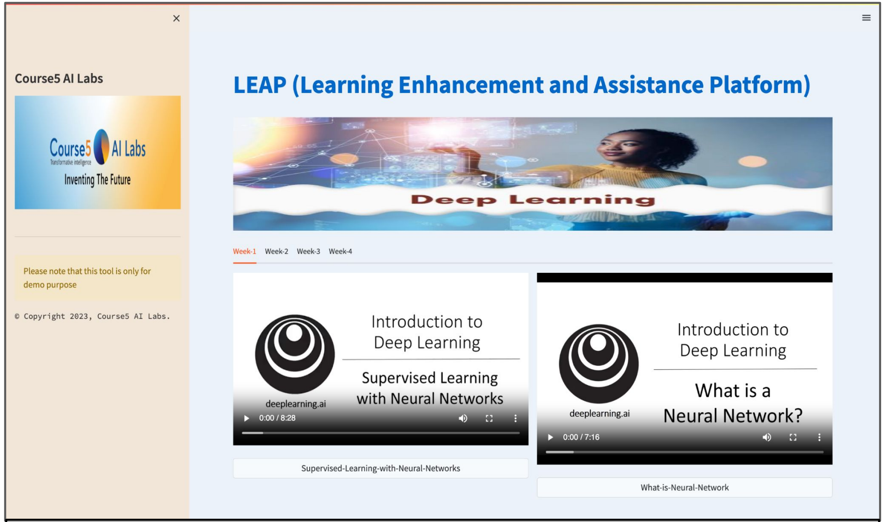
Algebra

Computer Vision

Deep Learning

Machine Learning

Screen-1 (Dashboard Page)



Course5 AI Labs

Please note that this tool is only for demo purpose

© Copyright 2023, Course5 AI Labs.

LEAP (Learning Enhancement and Assistance Platform)

Week-1 Week-2 Week-3 Week-4

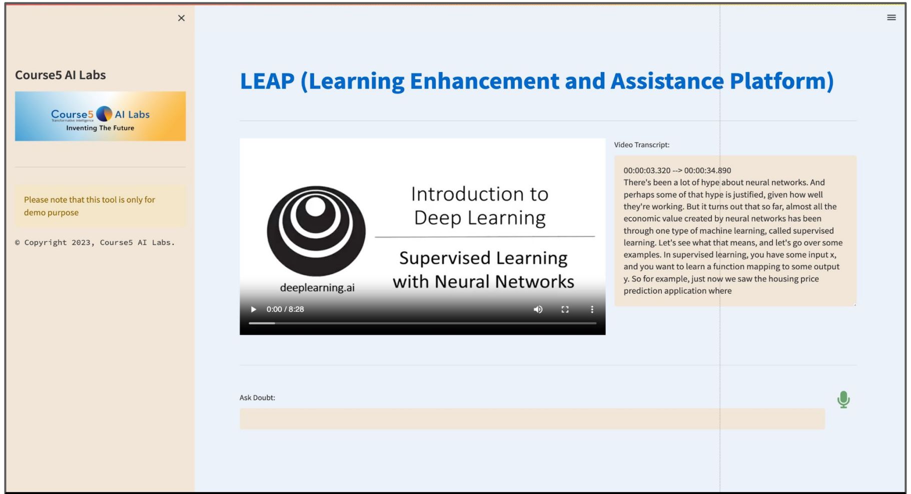
Supervised-Learning-with-Neural-Networks

What-is-Neural-Network

Course5 AI Labs
Inventing The Future
Deep Learning
Introduction to Deep Learning
Supervised Learning with Neural Networks
deeplearning.ai
0:00 / 8:28
Introduction to Deep Learning
What is a Neural Network?
deeplearning.ai
0:00 / 7:16

Screen-2 (Week wise Content Page)

Demo Screenshots



Course5 AI Labs

Course5 AI Labs
Inventing The Future

Please note that this tool is only for demo purpose

© Copyright 2023, Course5 AI Labs.

LEAP (Learning Enhancement and Assistance Platform)

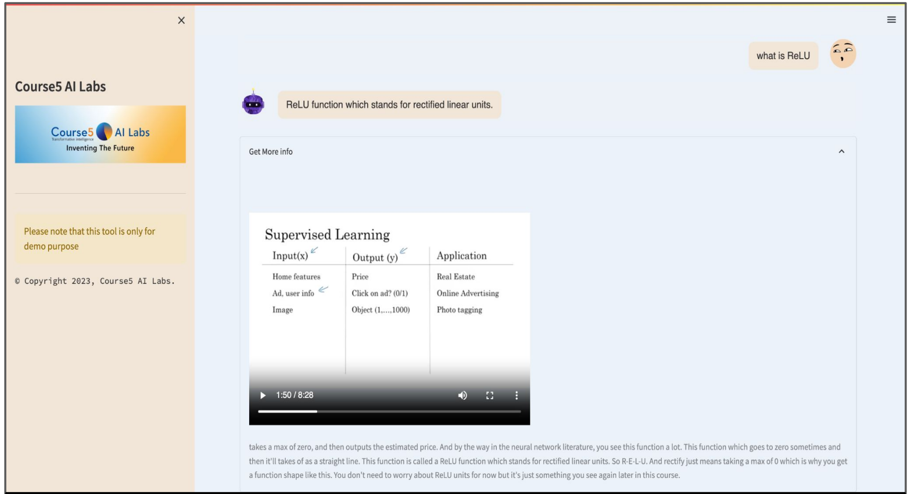
Introduction to Deep Learning
Supervised Learning with Neural Networks
deeplearning.ai
0:00 / 8:28

Video Transcript:

00:00:03.320 --> 00:00:34.890
There's been a lot of hype about neural networks. And perhaps some of that hype is justified, given how well they're working. But it turns out that so far, almost all the economic value created by neural networks has been through one type of machine learning, called supervised learning. Let's see what that means, and let's go over some examples. In supervised learning, you have some input x, and you want to learn a function mapping to some output y. So for example, just now we saw the housing price prediction application where

Ask Doubt:

Screen-3 (Ask Question/Doubt Page)



Course5 AI Labs

Please note that this tool is only for demo purpose

© Copyright 2023, Course5 AI Labs.

what is ReLU

ReLU function which stands for rectified linear units.

Get More info

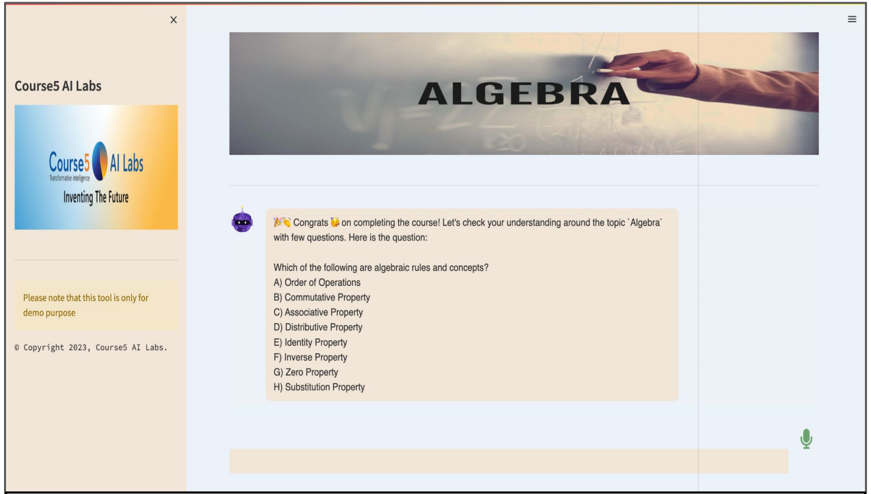
Supervised Learning

Input(x)	Output (y)	Application
Home features	Price	Real Estate
Ad, user info	Click on ad? (0/1)	Online Advertising
Image	Object (1,....,1000)	Photo tagging

takes a max of zero, and then outputs the estimated price. And by the way in the neural network literature, you see this function a lot. This function which goes to zero sometimes and then it'll takes of as a straight line. This function is called a ReLU function which stands for rectified linear units. So R-E-L-U. And rectify just means taking a max of 0 which is why you get a function shape like this. You don't need to worry about ReLU units for now but it's just something you see again later in this course.

Screen-4 (Ask Question/Doubt Page)

Demo Screenshots



Course5 AI Labs

Course5 AI Labs Transformative Intelligence Inventing The Future

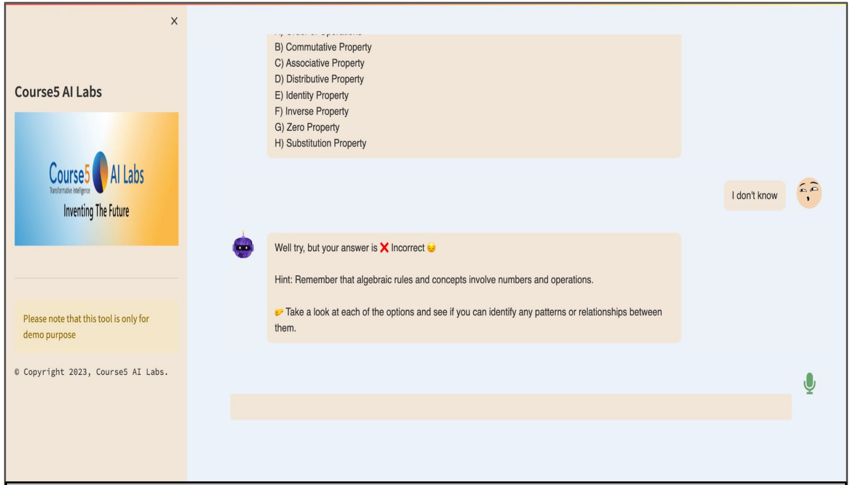
Please note that this tool is only for demo purpose

© Copyright 2023, Course5 AI Labs.

🎉👏 Congrats 🥳 on completing the course! Let's check your understanding around the topic `Algebra` with few questions. Here is the question:

Which of the following are algebraic rules and concepts?
A) Order of Operations
B) Commutative Property
C) Associative Property
D) Distributive Property
E) Identity Property
F) Inverse Property
G) Zero Property
H) Substitution Property

Screen-5 (Interactive Conversational AI Examiner Asks Question to student)



Course5 AI Labs

Course5 AI Labs Transformative Intelligence Inventing The Future

Please note that this tool is only for demo purpose

© Copyright 2023, Course5 AI Labs.

B) Commutative Property
C) Associative Property
D) Distributive Property
E) Identity Property
F) Inverse Property
G) Zero Property
H) Substitution Property

I don't know

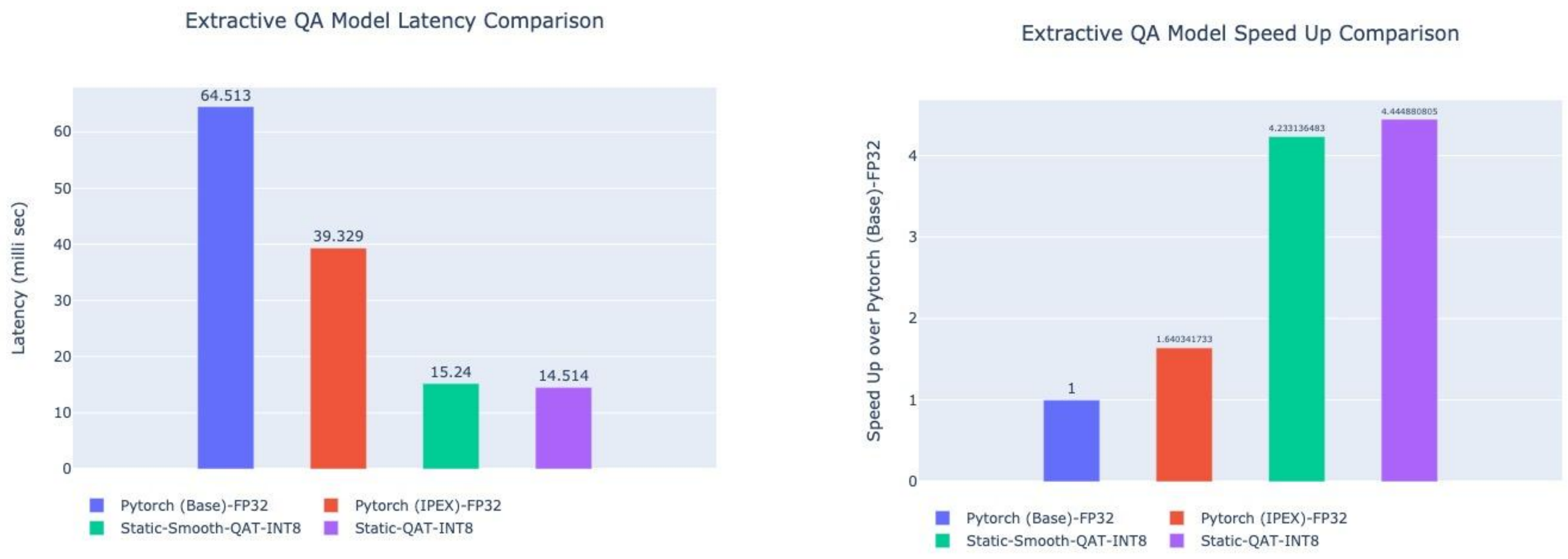
Well try, but your answer is ❌ Incorrect 😔

Hint: Remember that algebraic rules and concepts involve numbers and operations.

👉 Take a look at each of the options and see if you can identify any patterns or relationships between them.

Screen-6 (Interactive Conversational AI Examiner provides hints and motivates a student in case of a wrong answer)

Extractive QA Model (BERT Topology) Latency/Speed-Up Comparison with IPEX and Intel® Neural Compressor



Extractive QA Model Latency Comparison

Extractive QA Model Speed Up Comparison

*Fig: Latency/Speed-Up Benchmark result for **our Extractive Question Answering ALBERT Model (Multilingual)** on Intel® Dev Cloud machine (**Intel® Xeon® Platinum 8480+ (4th Gen: Sapphire Rapids) - 224v CPUs 503GB RAM**) with optimization using IPEX-FP32 and Static INT8-Quantization using Intel® Neural Compressor.*

For Ask Question/Doubt Extractive QA Model

Extractive QA Model (BERT Topology) Throughput/F1 Score Comparison with IPEX and Intel® Neural Compressor

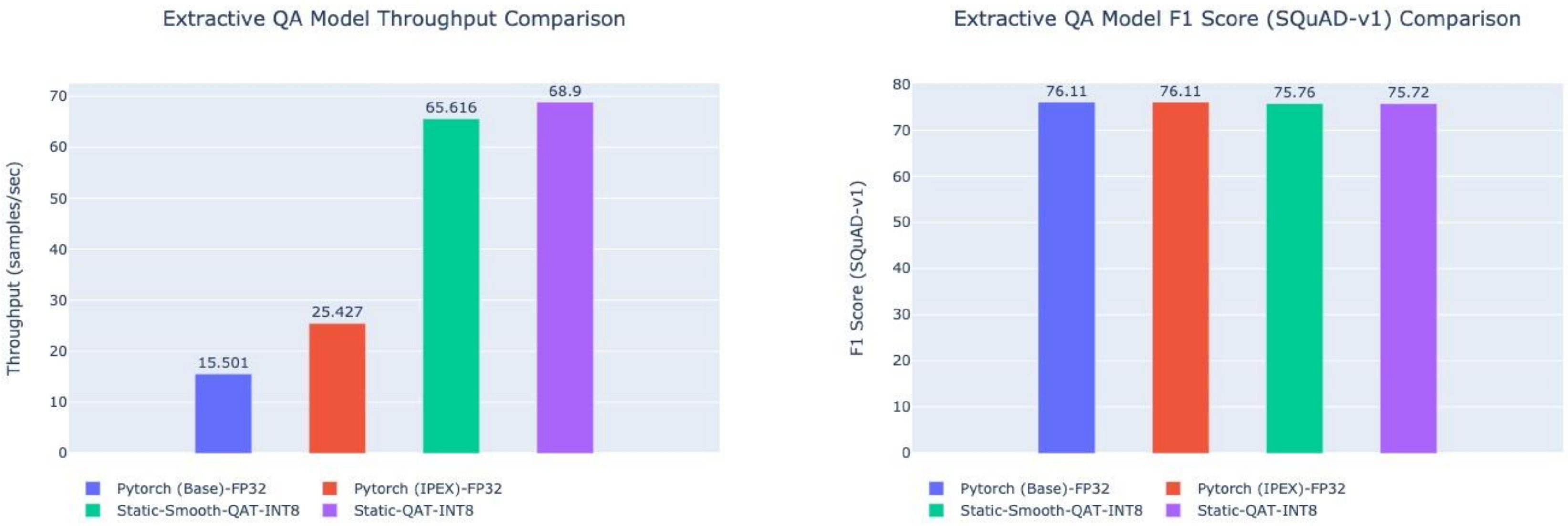


*Fig: Throughput/F1 Score Benchmark result for **our Extractive Question Answering ALBERT Model (Multilingual)** on Intel® Dev Cloud machine (**Intel® Xeon® Platinum 8480+ (4th Gen: Sapphire Rapids**) **- 224v CPUs 503GB RAM**) with optimization using IPEX-FP32 and Static INT8-Quantization using Intel® Neural Compressor. Also, the model (https://huggingface.co/ai4bharat/indic-bert) was fine-tuned on SQuAD-v1 dataset.*

For Ask Question/Doubt Extractive QA Model

Scikit-Learn (Base) vs Intel® Extension for Scikit-Learn

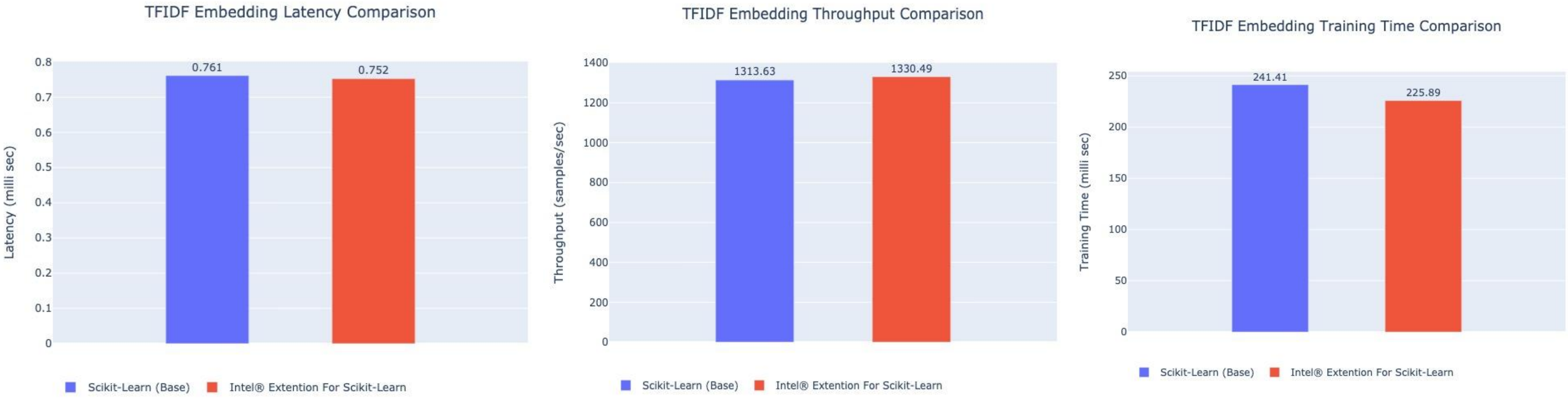
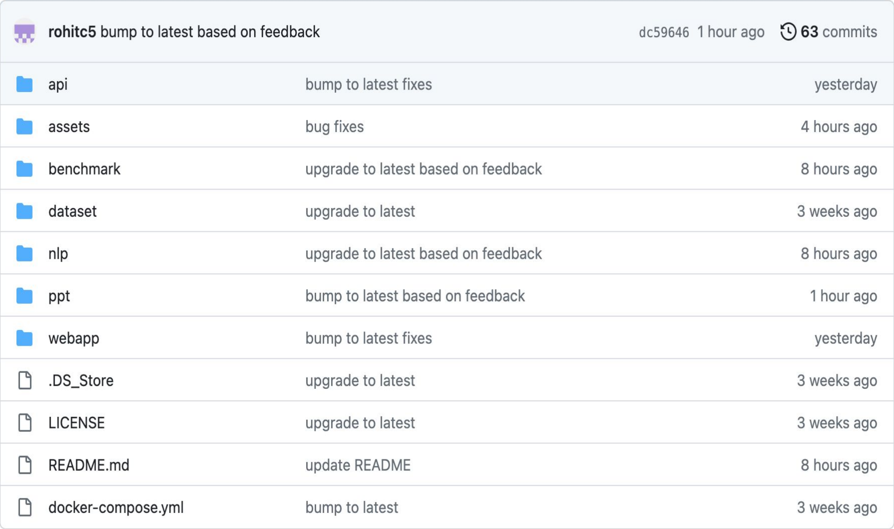


*Fig: Benchmark results for **TFIDFVectorizer** Embedding model during training and inference on Intel® Dev Cloud machine (**Intel® Xeon® Platinum 8480+ (4th Gen: Sapphire Rapids) - 224v CPUs 503GB RAM**). Please Note that we don't see much of a difference may be because we used a tiny dataset.*

For Ask Question/Doubt Embedding Model

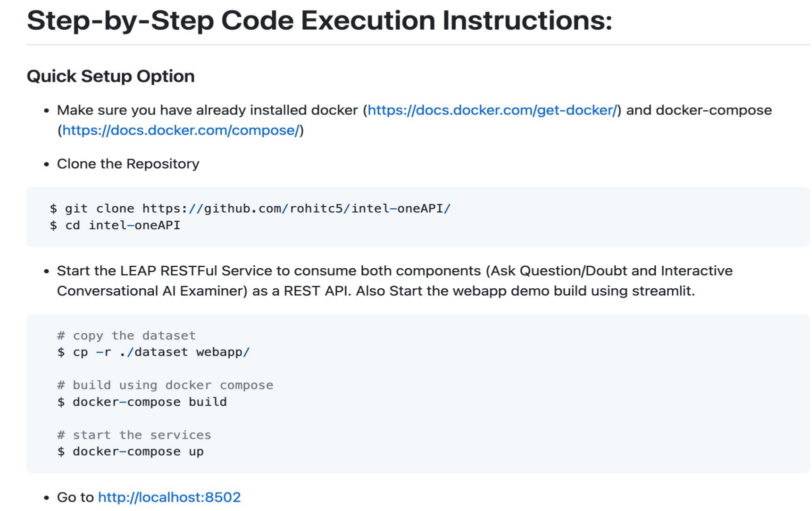
GitHub Link (Codes should be public and available after hackathon also)

https://github.com/rohitc5/intel-oneAPI



rohitc5 bump to latest based on feedback   dc59646 1 hour ago   63 commits

api	bump to latest fixes	yesterday
assets	bug fixes	4 hours ago
benchmark	upgrade to latest based on feedback	8 hours ago
dataset	upgrade to latest	3 weeks ago
nlp	upgrade to latest based on feedback	8 hours ago
ppt	bump to latest based on feedback	1 hour ago
webapp	bump to latest fixes	yesterday
.DS_Store	upgrade to latest	3 weeks ago
LICENSE	upgrade to latest	3 weeks ago
README.md	update README	8 hours ago
docker-compose.yml	bump to latest	3 weeks ago



Step-by-Step Code Execution Instructions:

Quick Setup Option

- Make sure you have already installed docker (https://docs.docker.com/get-docker/) and docker-compose (https://docs.docker.com/compose/)
- Clone the Repository

```
$ git clone https://github.com/rohitc5/intel-oneAPI/
$ cd intel-oneAPI
```

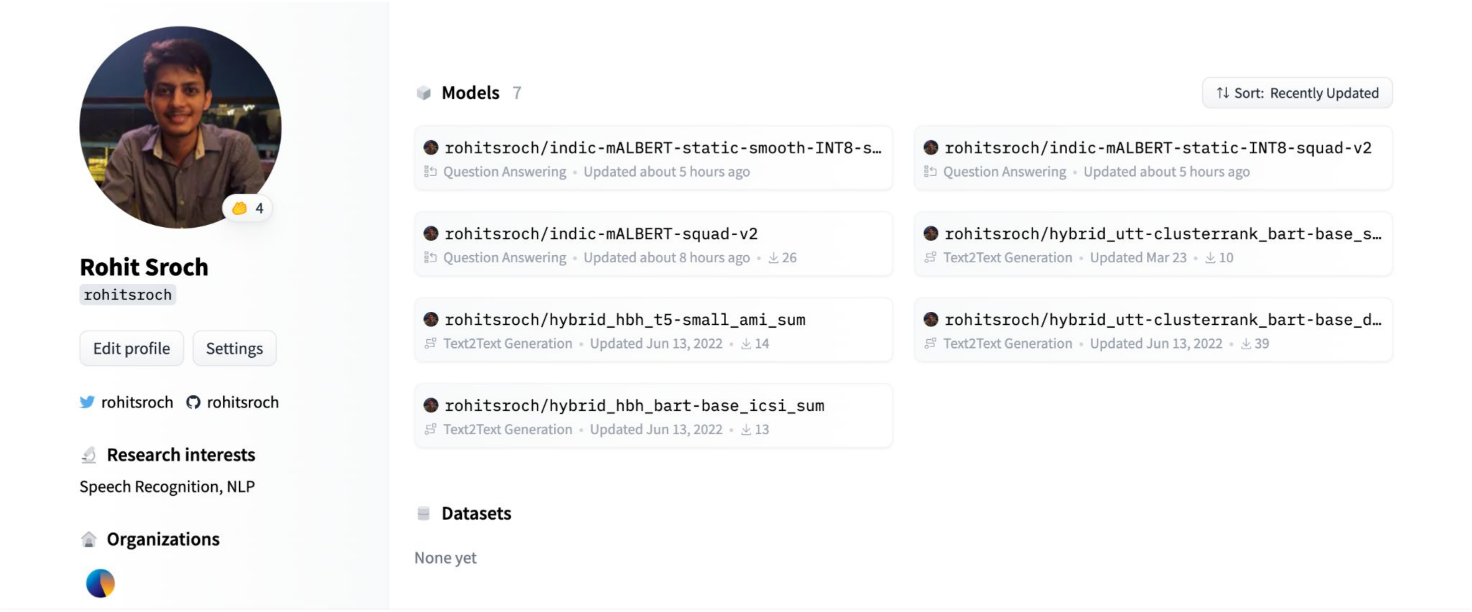
- Start the LEAP RESTFul Service to consume both components (Ask Question/Doubt and Interactive Conversational AI Examiner) as a REST API. Also Start the webapp demo build using streamlit.

```
# copy the dataset
$ cp -r ./dataset webapp/

# build using docker compose
$ docker-compose build

# start the services
$ docker-compose up
```

- Go to http://localhost:8502

Model Checkpoint Release

https://huggingface.co/rohitsroch



Rohit Sroch
rohitsroch

rohitsroch rohitsroch

Research interests

Speech Recognition, NLP

Organizations

Models 7

rohitsroch/indic-mALBERT-static-smooth-INT8-s...
Question Answering • Updated about 5 hours ago

rohitsroch/indic-mALBERT-static-INT8-squad-v2
Question Answering • Updated about 5 hours ago

rohitsroch/indic-mALBERT-squad-v2
Question Answering • Updated about 8 hours ago • 26

rohitsroch/hybrid_utt-clusterrank_bart-base_s...
Text2Text Generation • Updated Mar 23 • 10

rohitsroch/hybrid_hbh_t5-small_ami_sum
Text2Text Generation • Updated Jun 13, 2022 • 14

rohitsroch/hybrid_utt-clusterrank_bart-base_d...
Text2Text Generation • Updated Jun 13, 2022 • 39

rohitsroch/hybrid_hbh_bart-base_icsi_sum
Text2Text Generation • Updated Jun 13, 2022 • 13

Datasets

None yet

H2S
HACK2SKILL

intel.

oneAPI

<HACK>ATHON

THANK YOU