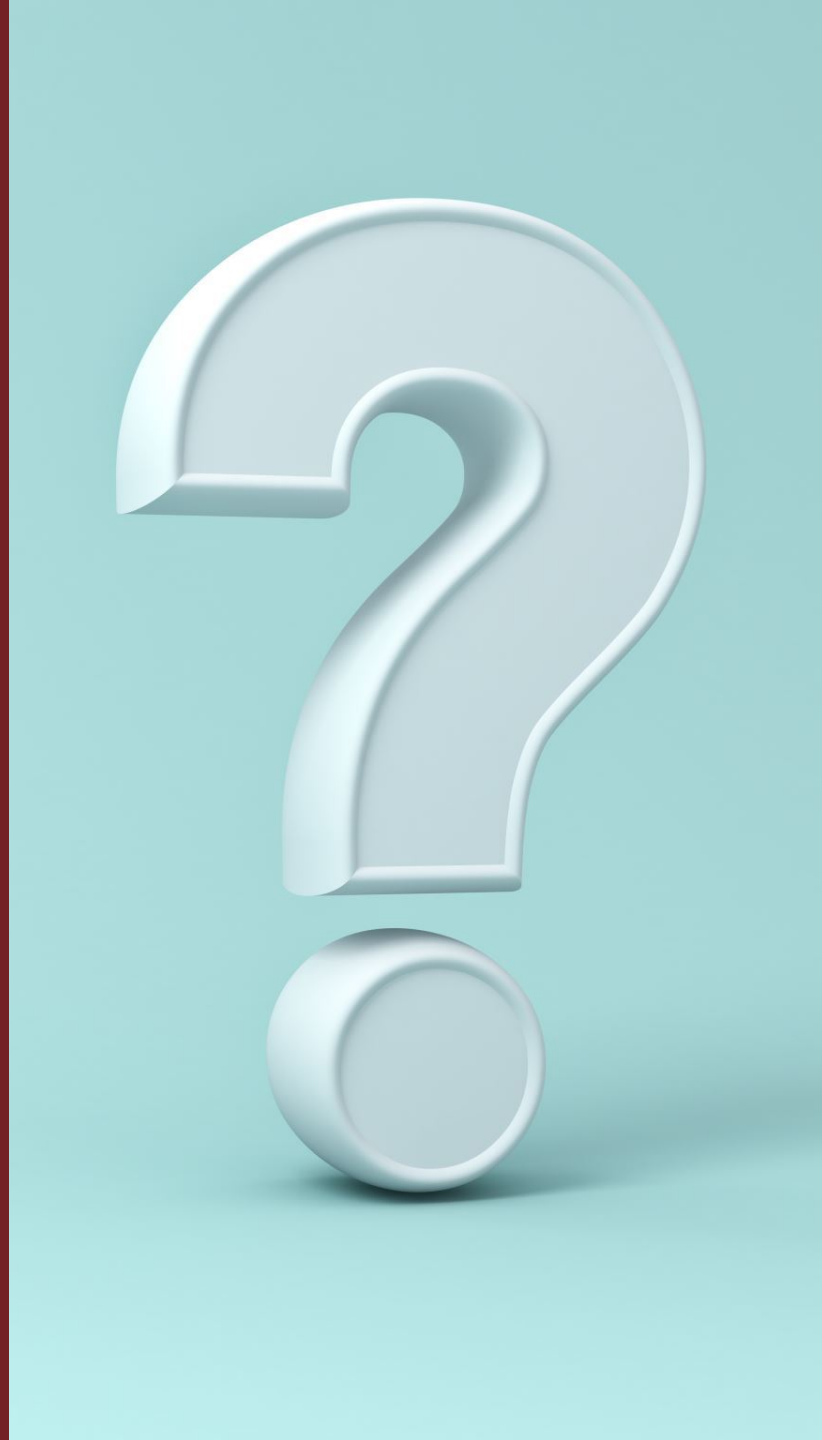
COMPARATIVE ANALYSIS: EVALUATING THE PERFORMANCE OF LARGE LANGUAGE MODELS ON VARIOUS NLP TASKS

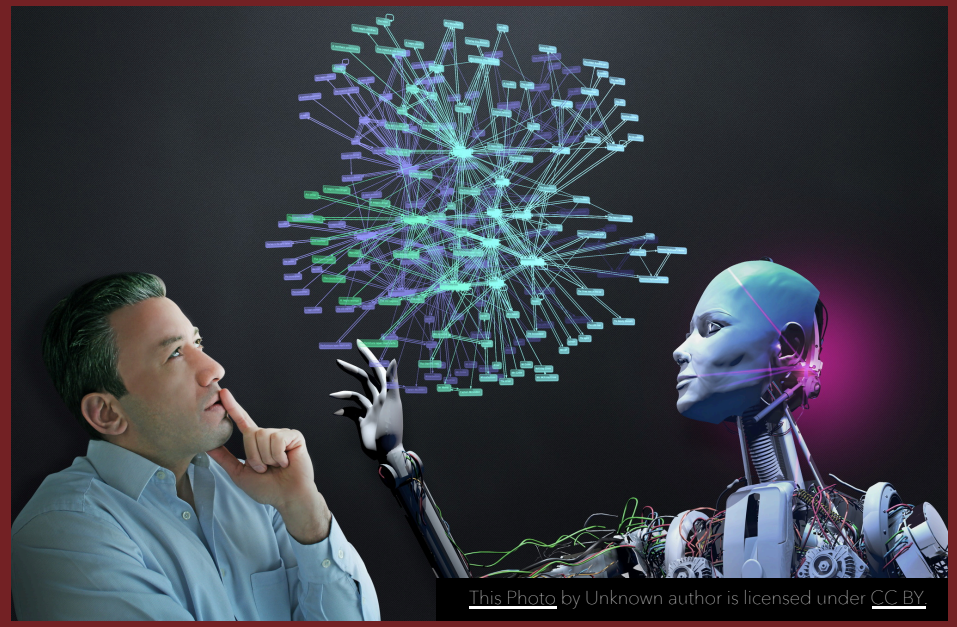
Progress Report - Abhinav, Arshia, Sai, Sushumna

PROBLEM STATEMENT

- **Evaluating large language models for in-context(zeroshot/one-shot/few-shot) performance on GLUE dataset :**
  - **CoLA** (Corpus of Linguistic Acceptability): *Binary* classification task to determine whether a given sentence is grammatically correct or not.
  - **SST-2** (Stanford Sentiment Treebank): *Binary* classification task to determine whether a given sentence has a positive or negative sentiment.
  - ***QNLI*** *(Question-answering Natural Language Inference): A natural language inference task to determine whether a given question can be answered by a given sentence*



BRIEF OVERVIEW: MODELS

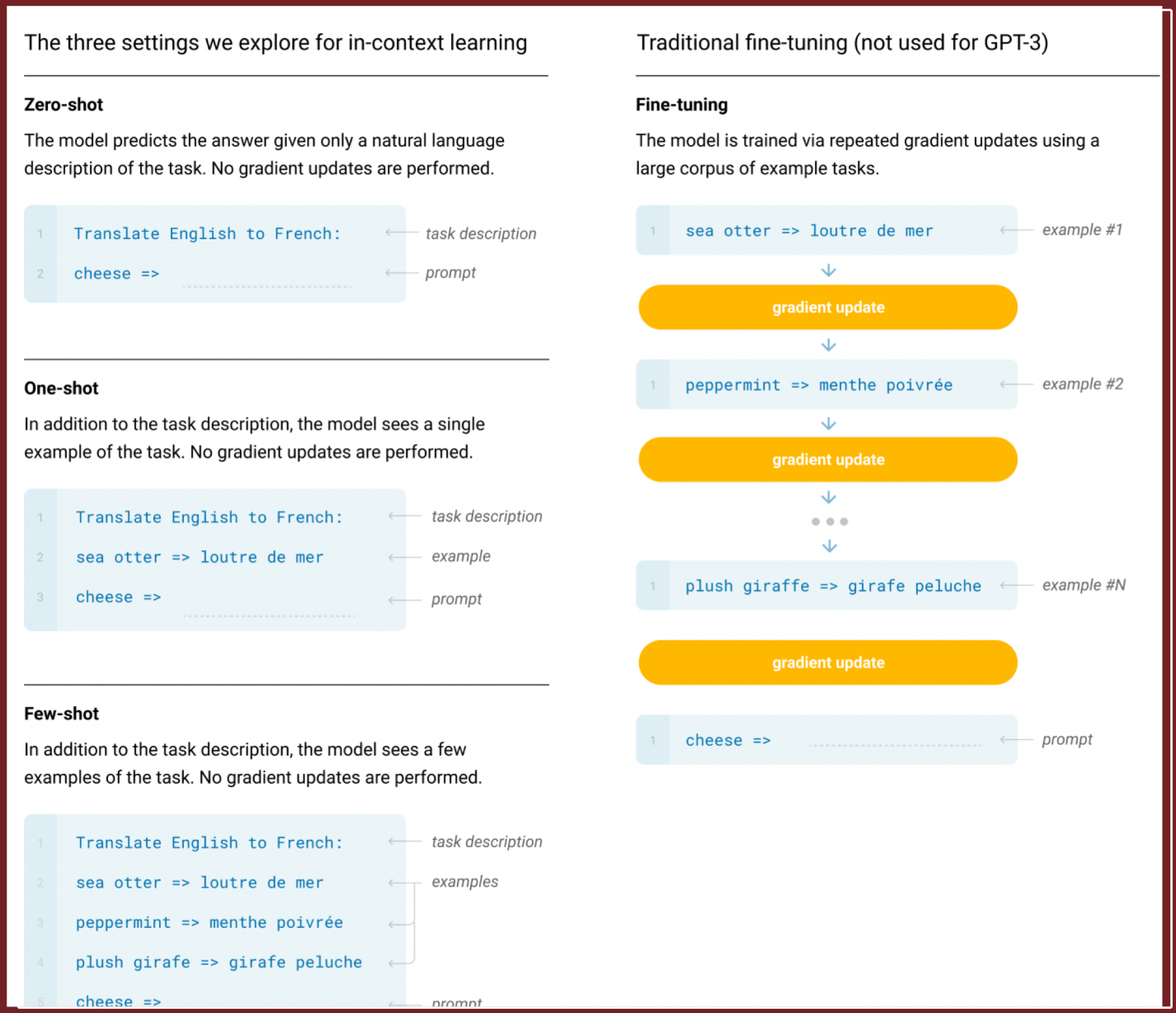


We selected Language models which were created as an open-source alternative to GPT-3.

1. **Bloom** : BigScience Large Open-science Open-access Multilingual Language Model by 1000 AI researchers
    - 176B params
    - 59 languages
2. ***LLaMA***: Large Language Model by Meta AI
    - *7B, 13B, 33B, and 65B parameters*
    - *trained on one trillion tokens*
3. **OPT 175B**: Open Pretrained Transformer by Meta AI
    - 175 billion parameters
4. **GPT Neo**: Generative Pre-trained Transformer-Neo formed by EleutherAI.
    - 125M, 350M, 1.3B, 2.7B parameters

FINE-TUNING AND IN-CONTEXT LEARNING

- Fine-tuning is a technique used to adapt pre-trained models to new data.
- In-context learning involves training models specifically for a particular context.



Problem Statement : Models, Tasks, Dataset

Shortlisted 4 models (LLaMA, GPT-NEO, OPT 175B, Bloom)

3 tasks (Grammatical Correctness, Sentiment Analysis, Q/A)

Zeroshot, One-shot, & Few-shot performance

LITERATURE REVIEW

- Previous research has focused on comparing different language models in the context of zero-shot, one-shot, and few-shot learning for various natural language processing tasks on the GLUE dataset.
- Some of the large language models that have been evaluated for these tasks include FLAN, GPT-3, and Bloom, all of which have shown promising performance.
- However, with the recent introduction of newer models such as Facebook's LlaMa, OPT 175B, GPT-Neo, and Bloom, there is a need for further evaluation of their performance in both zero-shot and fine-tuning scenarios.
- These Open-Source models equivalent to GPT 3 can potentially improve the state-of-the-art in natural language processing, especially in scenarios where labeled data is scarce.

BASELINE RESULTS GPT-NEO 125M ZEROSHOT EVALUATION

- We evaluated GPT-NEO 125M model on the GLUE sst2 dataset for the task of sentiment analysis using zero-shot classification technique.
- This model was never exposed to the GLUE dataset and was not finetuned

```
import torch
from transformers import AutoTokenizer, GPTNeoForSequenceClassification

tokenizer = AutoTokenizer.from_pretrained("EleutherAI/gpt-neo-125M")
model = GPTNeoForSequenceClassification.from_pretrained("EleutherAI/gpt-neo-125M")

with torch.no_grad():
    logits = model(**inputs).logits
```

BASELINE RESULTS: EVALUATION METRICS

- We calculated the accuracy, Precision, Recall, and F-1 score for the GPT-Neo model
- It seems like the GPT-Neo model is performing poorly on zero-shot task with an overall accuracy of 0.425.
- Further analysis is required for higher parameter models of GPT-Neo with respect to the above task.

```
zeroshot_eval(actual,predict)
```

```
Accuracy: 0.42545871559633025
Precision: [0.43803056 0.39929329]
Recall: [0.60280374 0.2545045 ]
F1-score: [0.50737463 0.310866657]
Weighted Precision: 0.41830653541001944
Weighted Recall: 0.42545871559633025
Weighted F1-score: 0.4073177769124113
```

STATE OF THE ART RESULTS

- We will compare the performance of our 4 Language Models by benchmarking their zero-shot capabilities against the most advanced industry models currently available for the specific tasks we have selected
- Source: https://gluebenchmark.com/leaderboard/

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Microsoft Alexander v-team	Turing ULR v6		91.3	73.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9	92.5	92.1	96.7	93.6	97.9	55.4
2	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7	92.4	97.9	51.4
3	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9	57.0
4	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7	93.2	96.6	53.3
5	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	51.7

CHALLENGES

- It seems that despite our best efforts, evaluating the Large Language Models using **Google Collab and powerful desktop systems has proven to be a challenging task**.
- We have encountered **compute limitations** that impeded our progress and hindered our ability to achieve the desired results.
- The immense amount of compute required for loading and running these models will require quite a powerful system so we have decided to use supercomputers like **Carbonate and Big Red 200 available at IU HPC**.

CHALLENGES: REQUIRED COMPUTE – FINETUNING BLOOM

```
trainer.train()
```

```
/usr/local/lib/python3.9/dist-packages/transformers/optimization.py:391: FutureW
  warnings.warn(
You're using a BloomTokenizerFast tokenizer. Please note that with a fast tokeni
```

[8699/12630 1:24:15 < 38:05, 1.72 it/s, Epoch 2.07/3]

Epoch	Training Loss	Validation Loss	Accuracy
1	0.352200	0.467875	0.821101
2	0.214600	0.470437	0.831422

[12456/12630 1:58:17 < 01:39, 1.75 it/s, Epoch 2.96/3]

Epoch	Training Loss	Validation Loss	Accuracy
1	0.352200	0.467875	0.821101
2	0.214600	0.470437	0.831422

Leads to...

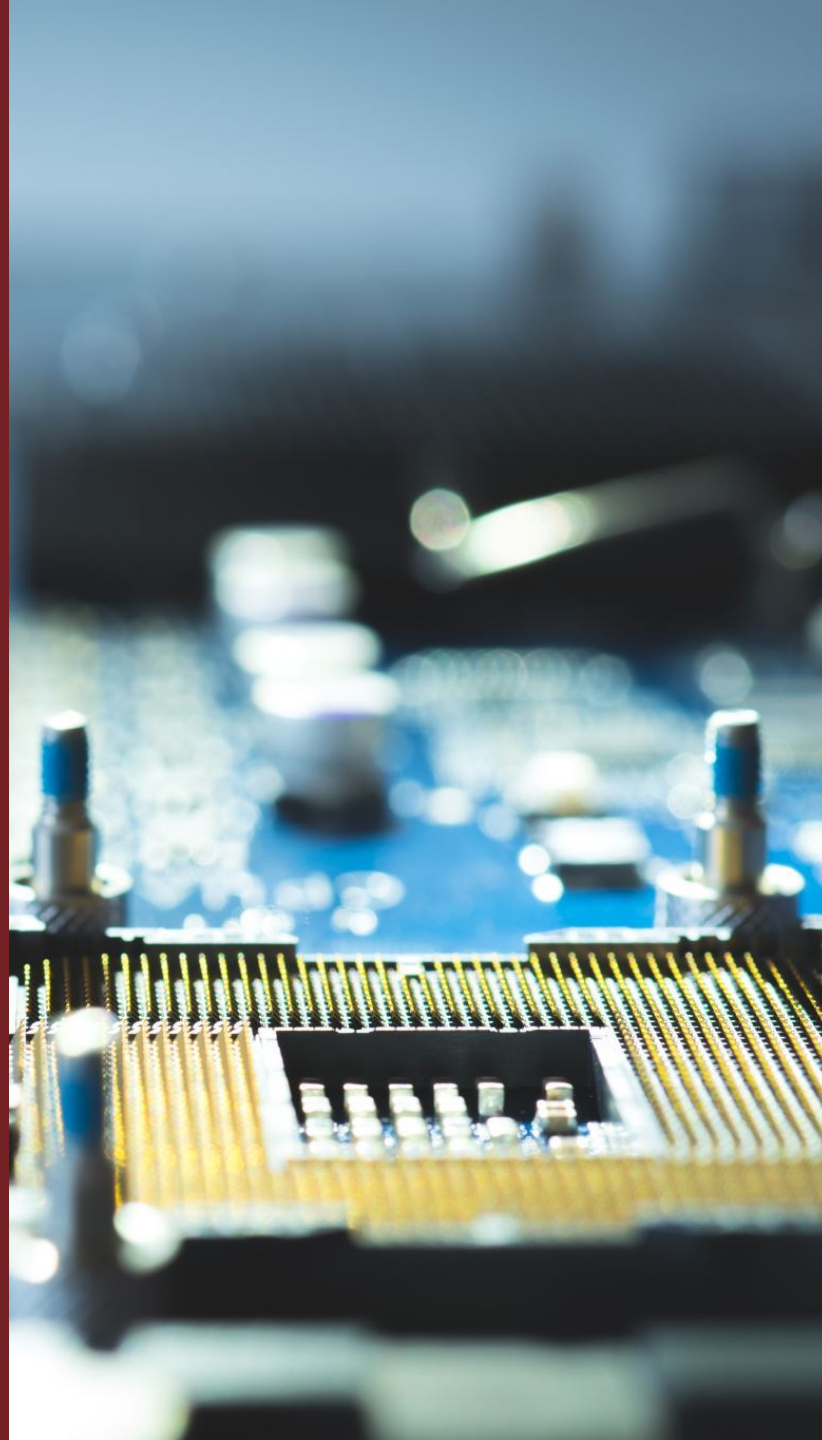
13.3-py3-none-any.whl (199 kB)

Cannot connect to GPU backend

You cannot currently connect to a GPU due to usage limits in Colab. Learn more

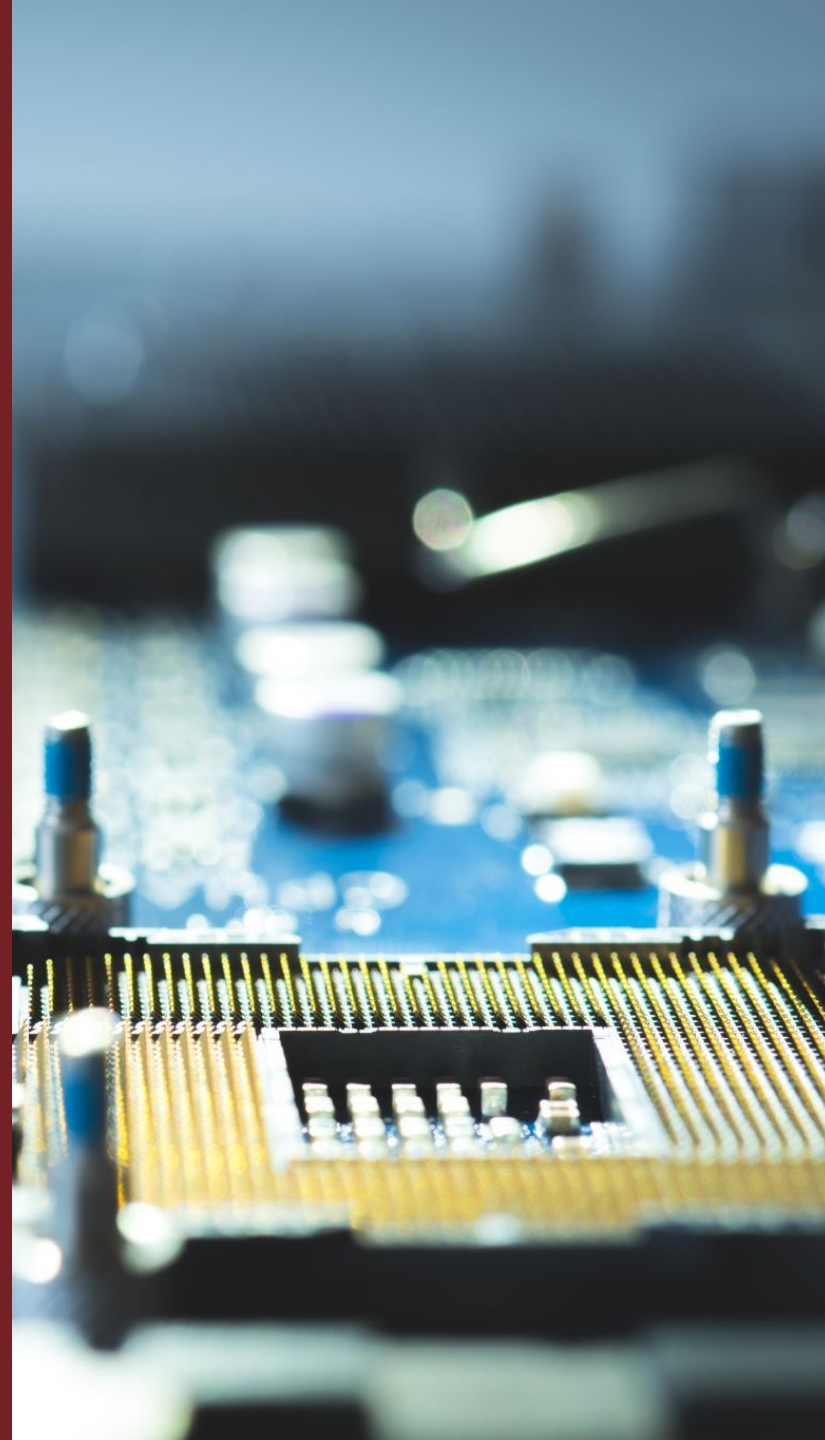
To get more access to GPUs, consider purchasing Colab compute units with Pay As You Go.

Close Connect without GPU



CHALLENGES: REQUIRED COMPUTE

- For GPT Neo 2. 7B – loading the model takes about 10GB of RAM.
- For OPT 175B – for loading and evaluation it takes about 40 GB of RAM.
- These models with the huge parameters will consume a huge amount of compute resources and if we try to finetune it will take even more amount of GPU's and CPU nodes.

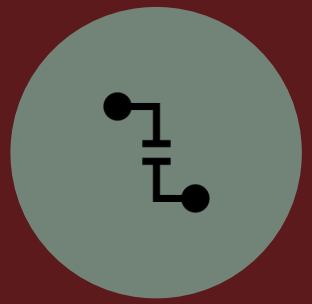


SOLUTION: COMPUTE

- Hugging Face provides Inference API which helps evaluate the models hosted on it.
- Carbonate: IU's large compute cluster which provides specialized deep learning (DL) and GPU partitions for researchers with deep learning applications and other applications that require GPUs.
  - 72 general-purpose compute nodes, each with 256 GB of RAM, and eight large-memory compute nodes, each with 512 GB of RAM.
  - *12 GPU-accelerated Lenovo ThinkSystem SD530 deep learning (DL) nodes, each equipped with two Intel Xeon Gold 6126 12-core CPUs, two NVIDIA GPU accelerators (eight with Tesla P100s; four with Tesla V100s), four 1.92 TB solid-state drives, and 192 GB of RAM.*
  - 24 GPU-accelerated Apollo 6500 nodes, each equipped with two Intel 6248 2.5 GHz 20-core CPUs, 768 GB of RAM, 4 NVIDIA V100-PCIE-32GB GPUs, and one 1.92 TB solid-state drive.

PROGRESS SUMMARY

Formulated the problem statement clearly and selected the **final models, tasks** to evaluate, and dataset.



Performed Zeroshot classification for GLUE SST2 with **GPT-Neo** model to get base results.

Finalised the **required Compute** resources needed for the task.

Obtained access and submitted first job to Carbonate for **evaluating GPT-NEO 125M** model.

PROGRESS SUMMARY : FURTHER WORK

Creation of RT project for resource allocation in carbonate.

Formulate all zeroshot classification tasks for LLaMA, OPT-175B, Bloom.

Evaluation for the specific tasks in GLUE(SST2, COLA, Q/A).

Compare Zeroshot/One-shot/Few-shot performance of models with different parameters using general evaluation metrics.

Try to Fine-Tune the specified models on GLUE dataset.

Compare the Fine-tuned and In-Context models.

Thank You!

Feel Free to ask any Questions