AI Tinkerers
Hackathon

Werecooked!

Fine-tuning Malaysian LLMs as Judges

bit.ly/werecookedai

Meet the team

werecooked is a team composed of Data Scientists weary of tabular data and SQL. Hence, when work makes us dull boys, we like to have fun with LLMs where things are slightly less predictable. While our team name may sound Pessimistic, rest assured that we're always cooking (as the gen Z's would say) by messing around with the latest AI trends and always learning to pick up the latest tools available

Naqib Asri

AI Engineer
Cloocus Sdn Bhd

Wan Adzhar Faiq

Data Scientist
Maybank

Ammar Azman

Data Scientist
Amanah Saham Nasional Berhad

Fahmi Zainal

Senior Data Scientist
INVOKE

Sukhdev Singh

Data Scientist
Maybank

Problem Statement

With the rise of AIs, LLMs have become an integral part of many industries.

However, in order to fully leverage these models, it isn’t just about deploying them. But it’s also about evaluating and refining hem to meet specific contextualised needs.

Traditional evaluation methods are outdated and ineffective at distinguishing high quality LLM responses from subpar ones.

While using human annotators gives great resuls, it requires significant effort and is difficult to scale across to multiple contexts.

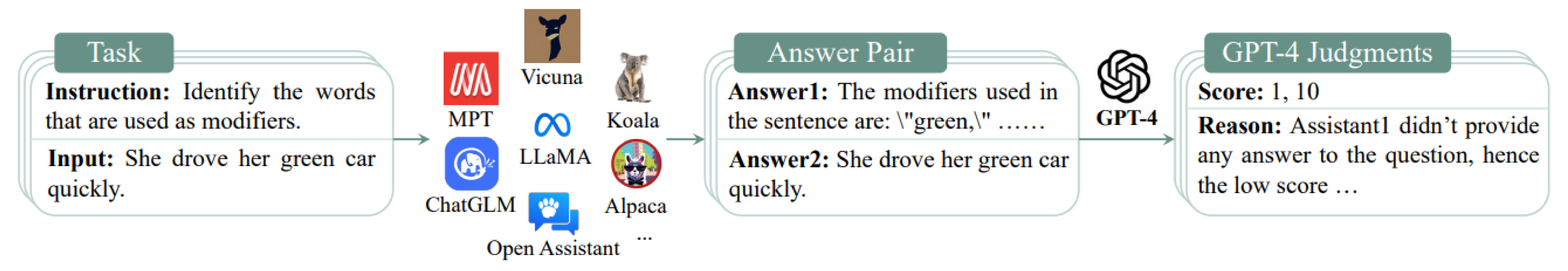
An emerging solution is using LLM Evaluators AKA “LLM-as-a-Judge” which uses LLMs to evaluae the quality of anoher LLM’s response.

For this Hackathon, we have developed an LLM Evaluator that judges the consistency of a statement based on an input document

Motivation

We are inspired by BAAIVision's work in employing LLMs as scalable judges for human preference. In our case, we're looking to finetune judges for predicting **logical/factual inconsistencies** as well as answering **yes/no** questions.

More crucially, this paper also presents promising **research where 7 Billion parameter models (which are much smaller) can approximate reasoning skills** close to if not better than **OpenAI's GPT4**. Let's see if we can replicate this!



Instruction: Identify the words that are used as modifiers.

Input: She drove her green car quickly.

Answer1: The modifiers used in the sentence are: \"green,\" ……

Answer2: She drove her green car quickly.

Score: 1, 10

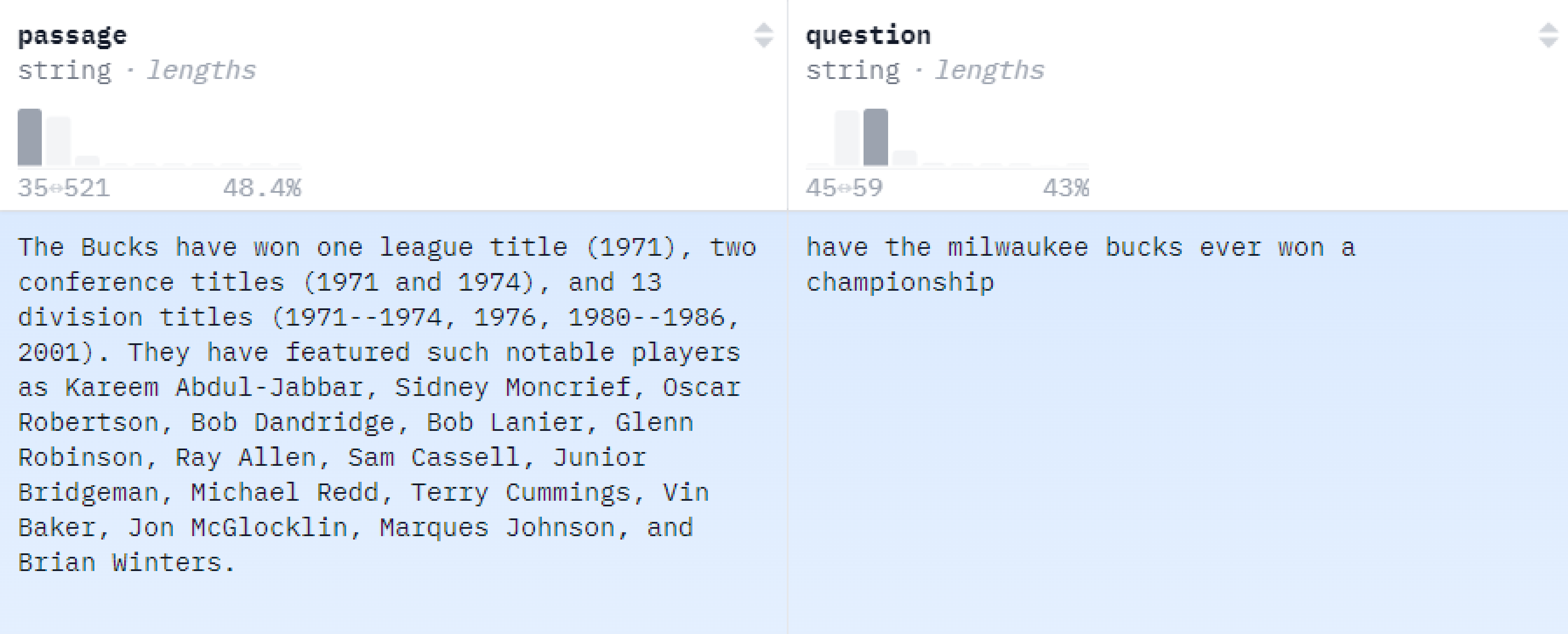
Reason: Assistant1 didn't provide any answer to the question, hence the low score …

Dataset Preparation

- We source our dataset from Google's Boolq a question answering dataset for yes/no questions containing 15942 examples; and a dataset from the FIB benchmark consisting of 3579 examples for evaluating the factual inconsistency of large language models.

- Both datasets are in English and useful for our use case in that they are binary. For boolq, we use it as the primary train/validation set due it being more.

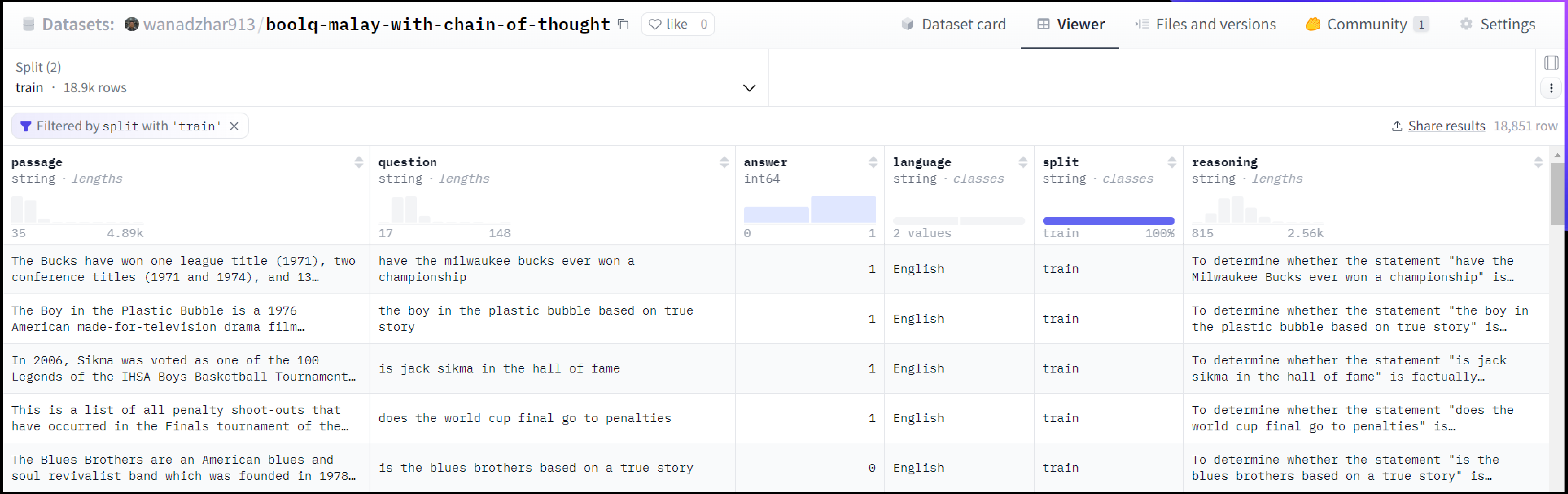
To ensure our model has high quality Malay examples, we employ OpenAI's 4o-mini to provide translation.



passage string · *lengths*	question string · *lengths*
35⇔521 48.4%	45⇔59 43%
The Bucks have won one league title (1971), two conference titles (1971 and 1974), and 13 division titles (1971--1974, 1976, 1980--1986, 2001). They have featured such notable players as Kareem Abdul-Jabbar, Sidney Moncrief, Oscar Robertson, Bob Dandridge, Bob Lanier, Glenn Robinson, Ray Allen, Sam Cassell, Junior Bridgeman, Michael Redd, Terry Cummings, Vin Baker, Jon McGlocklin, Marques Johnson, and Brian Winters.	have the milwaukee bucks ever won a championship

Sample 1 : Sample of a question-answer pair from Google's Boolq

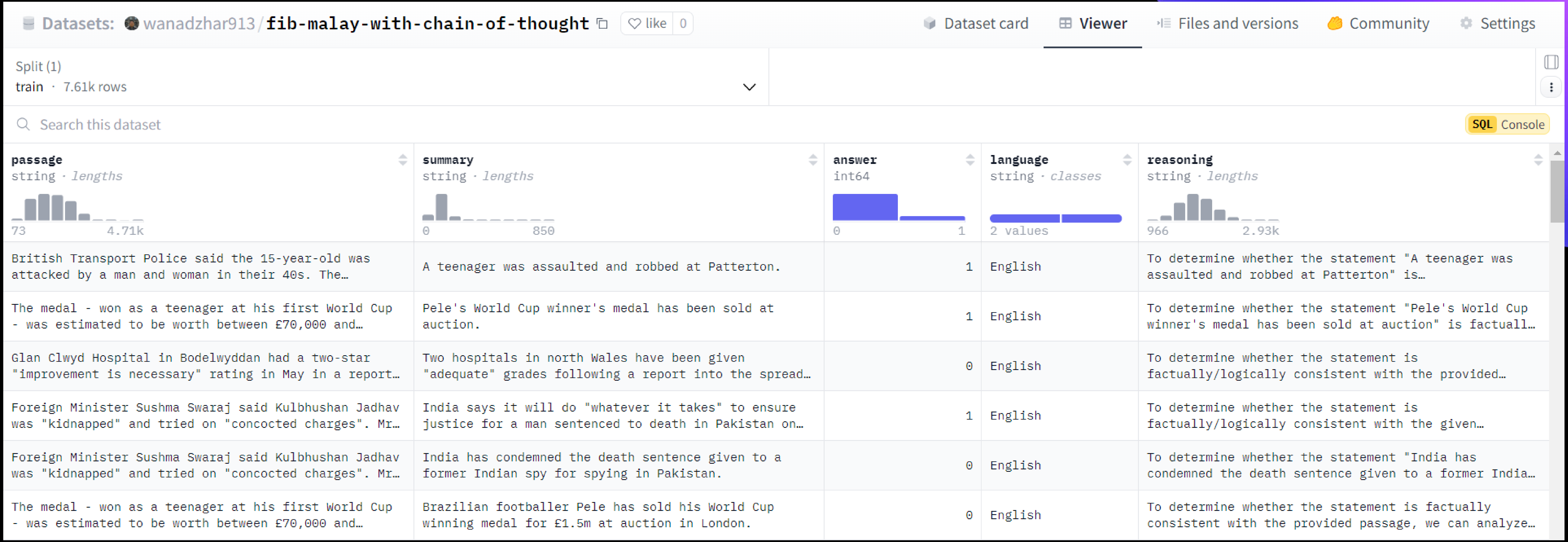
Dataset Preparation



passage string · *lengths* 35 4.89k	question string · *lengths* 17 148	answer int64 0 1	language string · *classes* 2 values	split string · *classes* train 100%	reasoning string · *lengths* 815 2.56k
The Bucks have won one league title (1971), two conference titles (1971 and 1974), and 13…	have the milwaukee bucks ever won a championship	1	English	train	To determine whether the statement "have the Milwaukee Bucks ever won a championship" is…
The Boy in the Plastic Bubble is a 1976 American made-for-television drama film…	the boy in the plastic bubble based on true story	1	English	train	To determine whether the statement "the boy in the plastic bubble based on true story" is…
In 2006, Sikma was voted as one of the 100 Legends of the IHSA Boys Basketball Tournament…	is jack sikma in the hall of fame	1	English	train	To determine whether the statement "is jack sikma in the hall of fame" is factually…
This is a list of all penalty shoot-outs that have occurred in the Finals tournament of the…	does the world cup final go to penalties	1	English	train	To determine whether the statement "does the world cup final go to penalties" is…
The Blues Brothers are an American blues and soul revivalist band which was founded in 1978…	is the blues brothers based on a true story	0	English	train	To determine whether the statement "is the blues brothers based on a true story" is…

Sample 2 : https://huggingface.co/datasets/wanadzhar913/boolq-malay-with-chain-of-thought

Dataset Preparation



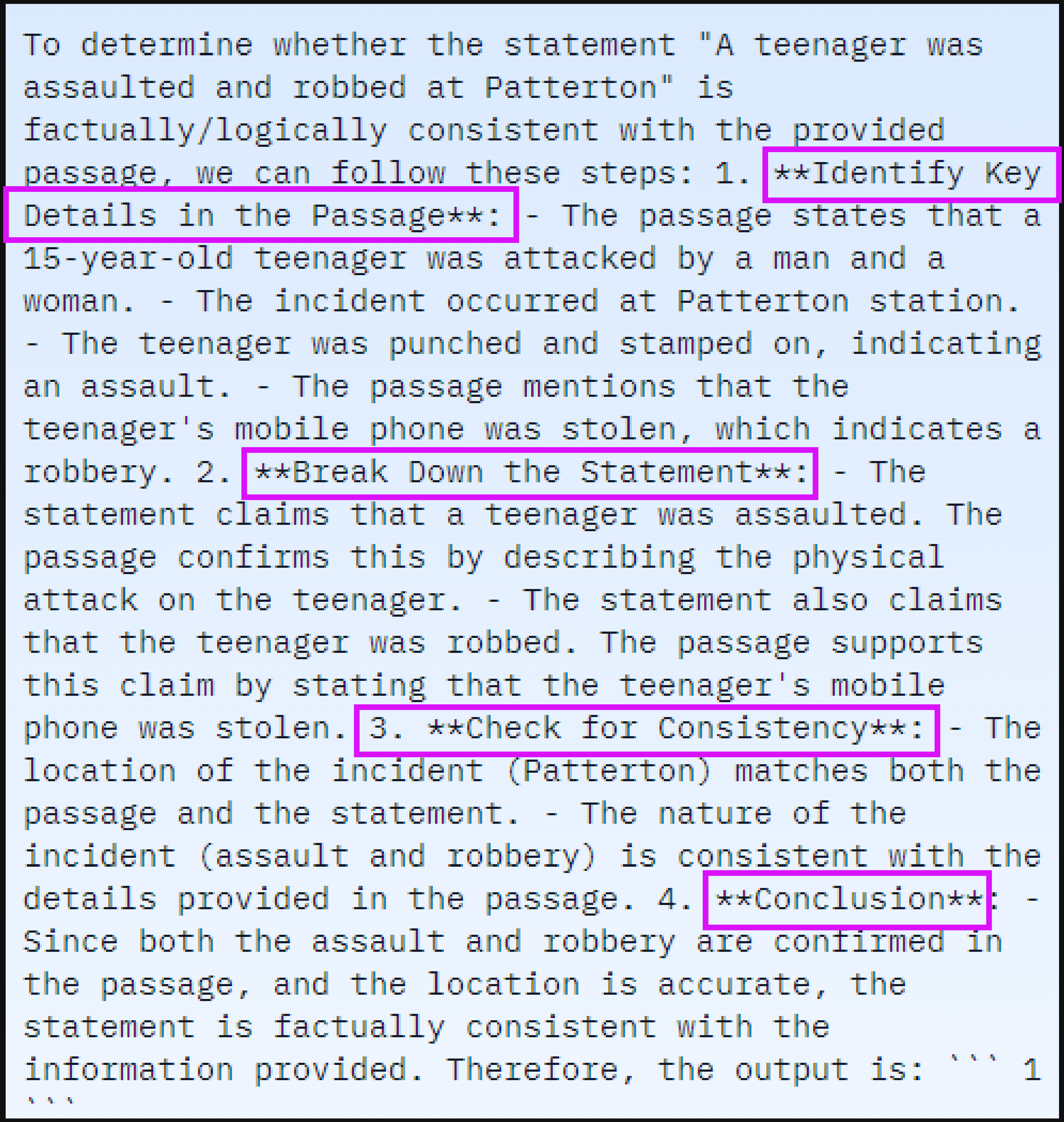
passage string · lengths 73 – 4.71k	summary string · lengths 0 – 850	answer int64 0 – 1	language string · classes 2 values	reasoning string · lengths 966 – 2.93k
British Transport Police said the 15-year-old was attacked by a man and woman in their 40s. The…	A teenager was assaulted and robbed at Patterton.	1	English	To determine whether the statement "A teenager was assaulted and robbed at Patterton" is…
The medal - won as a teenager at his first World Cup - was estimated to be worth between £70,000 and…	Pele's World Cup winner's medal has been sold at auction.	1	English	To determine whether the statement "Pele's World Cup winner's medal has been sold at auction" is factuall…
Glan Clwyd Hospital in Bodelwyddan had a two-star "improvement is necessary" rating in May in a report…	Two hospitals in north Wales have been given "adequate" grades following a report into the spread…	0	English	To determine whether the statement is factually/logically consistent with the provided…
Foreign Minister Sushma Swaraj said Kulbhushan Jadhav was "kidnapped" and tried on "concocted charges". Mr…	India says it will do "whatever it takes" to ensure justice for a man sentenced to death in Pakistan on…	1	English	To determine whether the statement is factually/logically consistent with the given…
Foreign Minister Sushma Swaraj said Kulbhushan Jadhav was "kidnapped" and tried on "concocted charges". Mr…	India has condemned the death sentence given to a former Indian spy for spying in Pakistan.	0	English	To determine whether the statement "India has condemned the death sentence given to a former India…
The medal - won as a teenager at his first World Cup - was estimated to be worth between £70,000 and…	Brazilian footballer Pele has sold his World Cup winning medal for £1.5m at auction in London.	0	English	To determine whether the statement is factually consistent with the provided passage, we can analyze…

Sample 3 : https://huggingface.co/datasets/wanadzhar913/fib-malay-with-chain-of-thought

Dataset Preparation

We also generate a ‘reasoning’ column to enable smaller models to perform complex reasoning tasks using OpenAI’s 4o-mini.

Zhang et. al found that small models under Fine-tune-CoT can even outperform their very large teachers in some tasks. With diverse reasoning, the performance of Fine-tune-CoT was highly scalable and led to high sample efficiency and notable reasoning performance even with few-shot training examples.

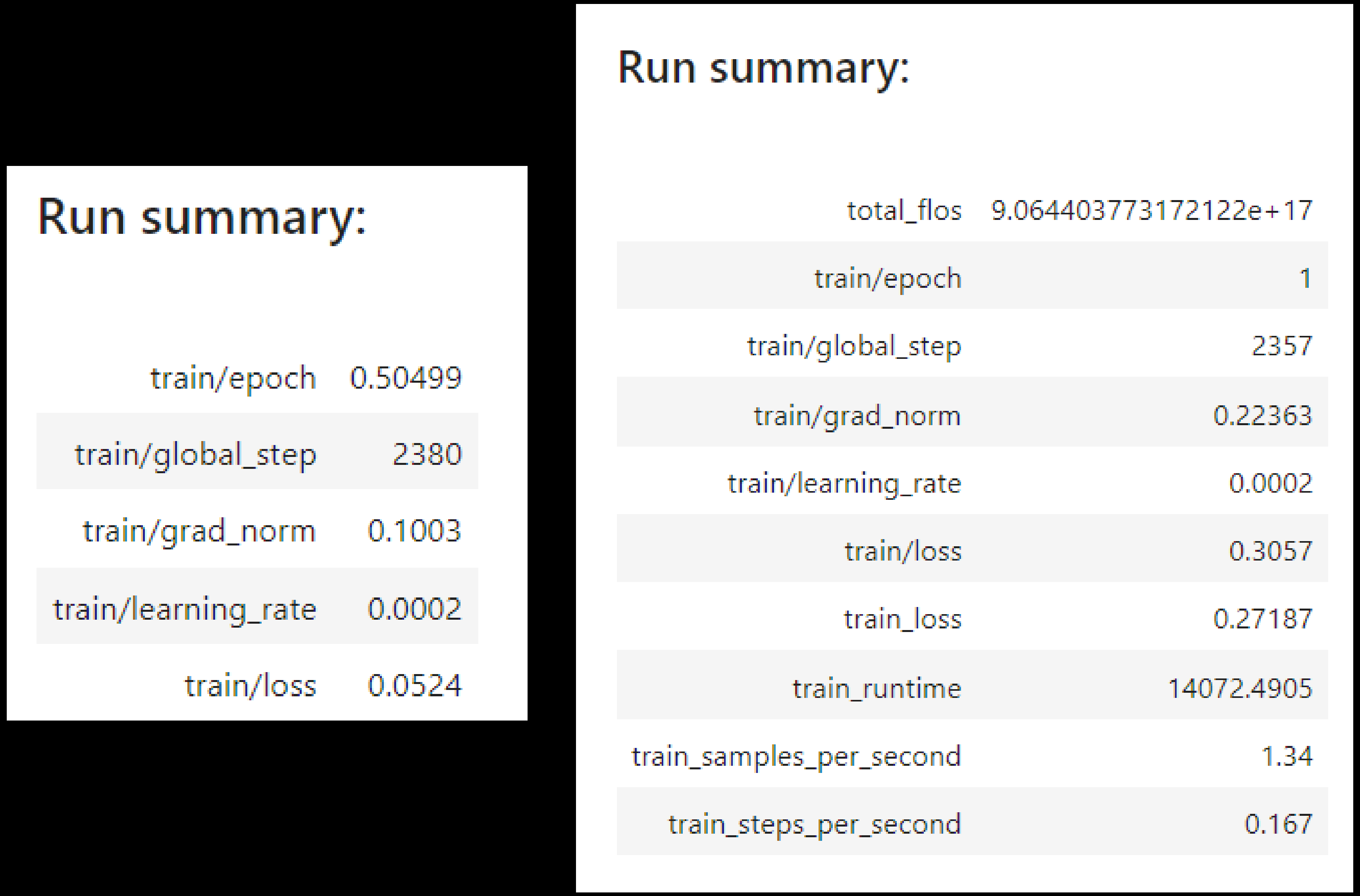


To determine whether the statement "A teenager was assaulted and robbed at Patterton" is factually/logically consistent with the provided passage, we can follow these steps: 1. **Identify Key Details in the Passage**: - The passage states that a 15-year-old teenager was attacked by a man and a woman. - The incident occurred at Patterton station. - The teenager was punched and stamped on, indicating an assault. - The passage mentions that the teenager's mobile phone was stolen, which indicates a robbery. 2. **Break Down the Statement**: - The statement claims that a teenager was assaulted. The passage confirms this by describing the physical attack on the teenager. - The statement also claims that the teenager was robbed. The passage supports this claim by stating that the teenager's mobile phone was stolen. 3. **Check for Consistency**: - The location of the incident (Patterton) matches both the passage and the statement. - The nature of the incident (assault and robbery) is consistent with the details provided in the passage. 4. **Conclusion**: - Since both the assault and robbery are confirmed in the passage, and the location is accurate, the statement is factually consistent with the information provided. Therefore, the output is: ``` 1 ```

Sample 4: Sample of Chain-of-Thought Reasoning generated using OpenAI’s 4o-mini

Finetuning

- We employ Huggingface’s trl (Transformer Reinforcement Learning) for Supervised Fine-tuning and finetune Mesolitica’s https://huggingface.co/mesolitica/malaysian-mistral-7b-32k-instructions-v4 model
- We train 2 models, one primarily focused on classification (V2) and the other on both classification and reasoning (V3).
- Both models were trained on the following hyper-parameters. Training time/Epochs was intentionally different between the 2 models due to resource constraints (Google Colab compute credits). Due to V2 reaching a lower loss early, we decided to end it’s training prematurely at 0.51 epochs, while V3 received 1 epoch’s worth of training.
- The Weights & Biases run for both models can be found here for V2 & V3.
- A small batch size was used to avoid **OutOfMemoryErrors** although in the future, the **gradient_accumulation_steps** parameter should be increased as well.



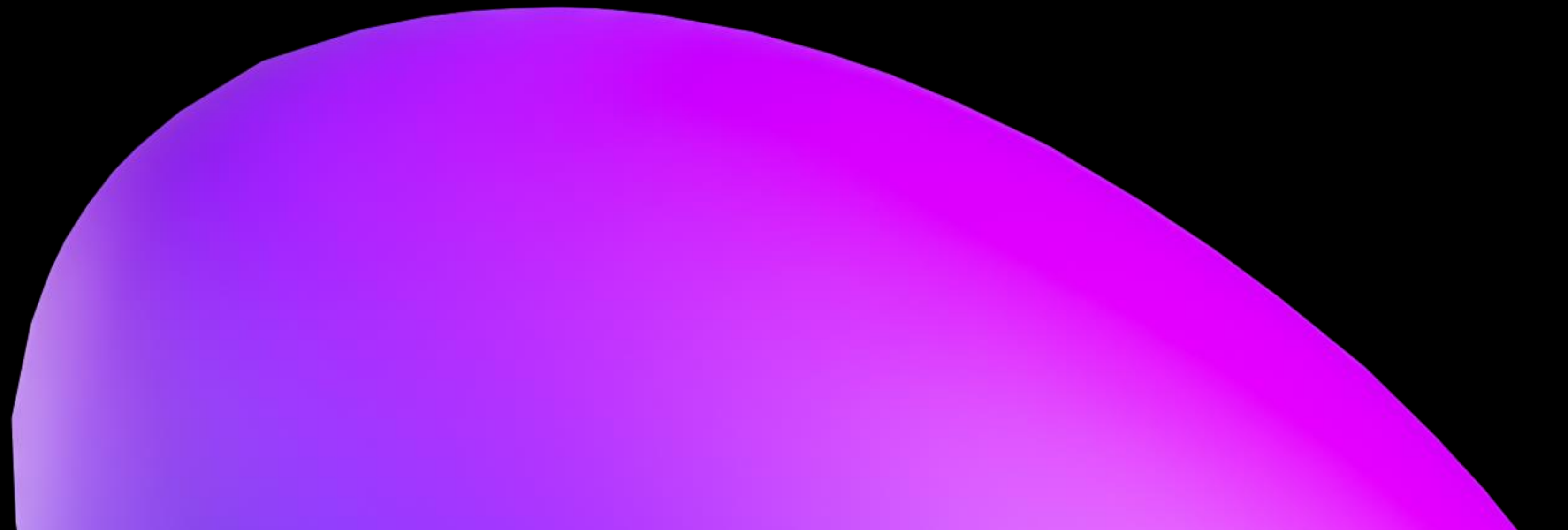
Run summary:

train/epoch	0.50499
train/global_step	2380
train/grad_norm	0.1003
train/learning_rate	0.0002
train/loss	0.0524

Run summary:

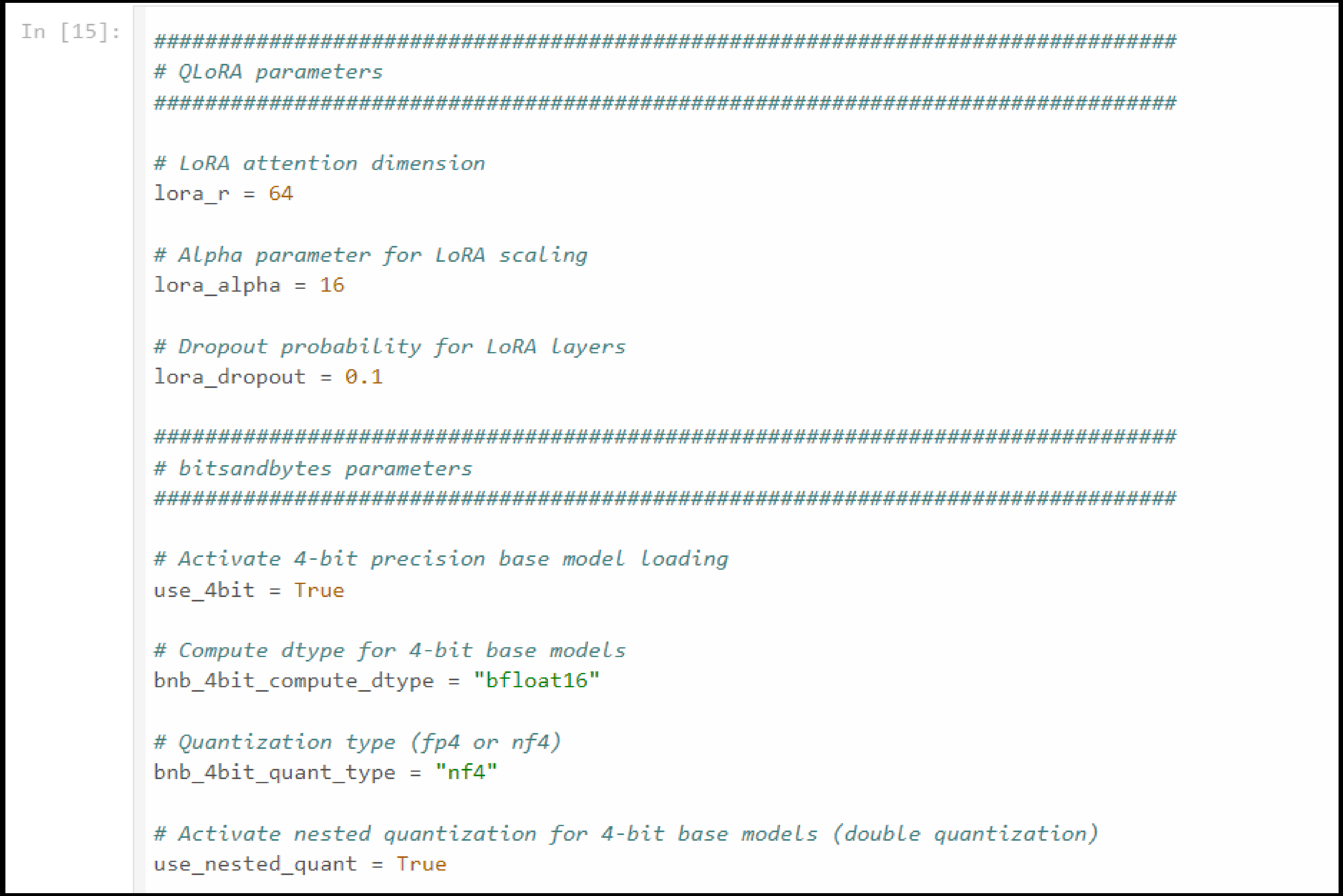
total_flos	9.064403773172122e+17
train/epoch	1
train/global_step	2357
train/grad_norm	0.22363
train/learning_rate	0.0002
train/loss	0.3057
train_loss	0.27187
train_runtime	14072.4905
train_samples_per_second	1.34
train_steps_per_second	0.167

Fig. 2 : The left is the training outcome for V2 and the right is for V3.



Finetuning

- Very crucially, we load our model in 4-bit precision using BitsandBytes & train our model using QLoRA (with a Rank of 64) to minimize memory usage during training.



In [15]:

```
################################################################################
# QLoRA parameters
################################################################################

# LoRA attention dimension
lora_r = 64

# Alpha parameter for LoRA scaling
lora_alpha = 16

# Dropout probability for LoRA layers
lora_dropout = 0.1

################################################################################
# bitsandbytes parameters
################################################################################

# Activate 4-bit precision base model loading
use_4bit = True

# Compute dtype for 4-bit base models
bnb_4bit_compute_dtype = "bfloat16"

# Quantization type (fp4 or nf4)
bnb_4bit_quant_type = "nf4"

# Activate nested quantization for 4-bit base models (double quantization)
use_nested_quant = True
```

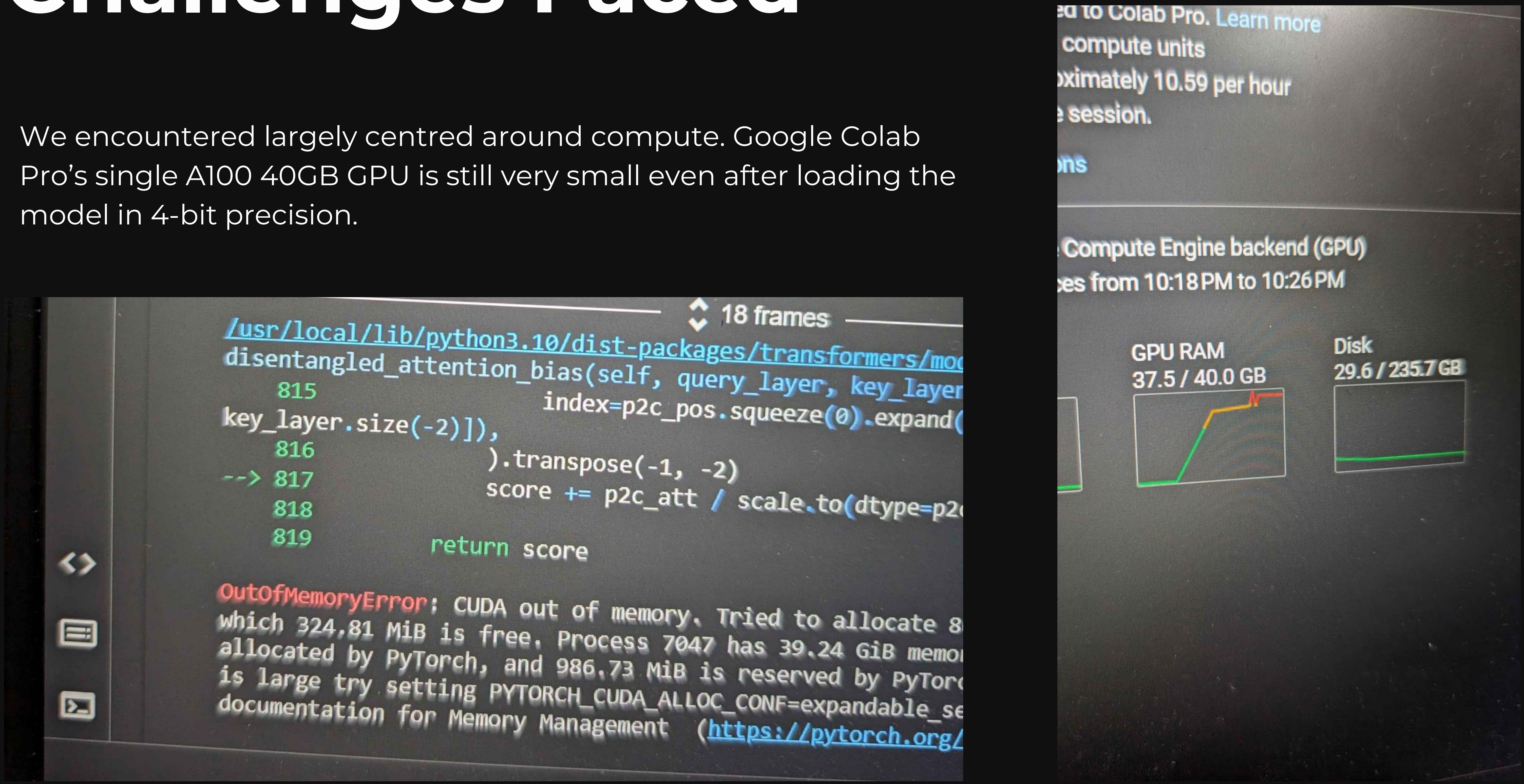
Results

	OpenAI 4o-mini	Finetuned-Malaysian-Mistral-V2	Finetuned-Malaysian-Mistral-V3*
F1	80%	74%	69%
Accuracy	78%	65%	61%
Precision	83%	65%	69%
Recall	77%	85%	70%

While we did see better performance with V3 on selected examples, the overall score suffered due to V3's inability to consistently return JSON responses.
*****The evaluation dataset was constructed by taking the first 200 samples from the evaluation set from Boolq-Malay & FIB-Malay dataset.***

Challenges Faced

We encountered largely centred around compute. Google Colab Pro’s single A100 40GB GPU is still very small even after loading the model in 4-bit precision.



```
18 frames
/usr/local/lib/python3.10/dist-packages/transformers/moc
disentangled_attention_bias(self, query_layer, key_layer
    815                      index=p2c_pos.squeeze(0).expand(
key_layer.size(-2)]),
    816                  ).transpose(-1, -2)
--> 817                  score += p2c_att / scale.to(dtype=p2c
    818
    819              return score

OutOfMemoryError: CUDA out of memory. Tried to allocate 8
which 324.81 MiB is free. Process 7047 has 39.24 GiB memor
allocated by PyTorch, and 986.73 MiB is reserved by PyTorc
is large try setting PYTORCH_CUDA_ALLOC_CONF=expandable_se
documentation for Memory Management  (https://pytorch.org/
```

ed to Colab Pro. Learn more
compute units
oximately 10.59 per hour
e session.
ons

Compute Engine backend (GPU)
ces from 10:18 PM to 10:26 PM

GPU RAM
37.5 / 40.0 GB

Disk
29.6 / 235.7 GB

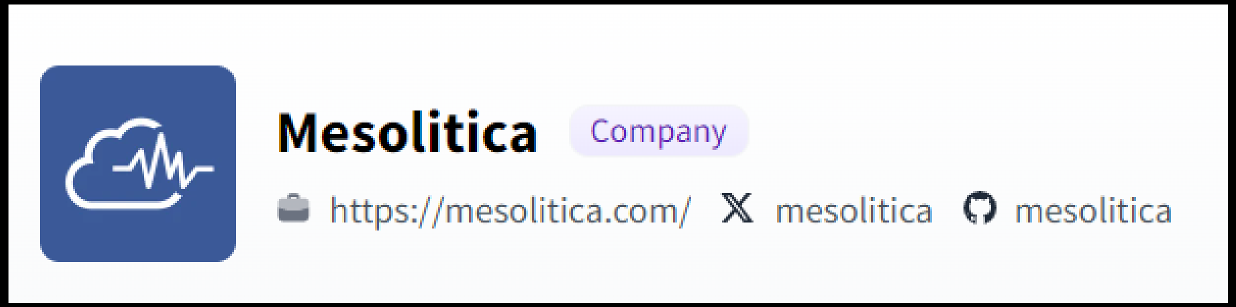
Fig. 3 : You can already imagine the horror on our faces as we saw these prints

Thank you!

Joseph from DocuAsk too for the guidance and OpenAI credits :)

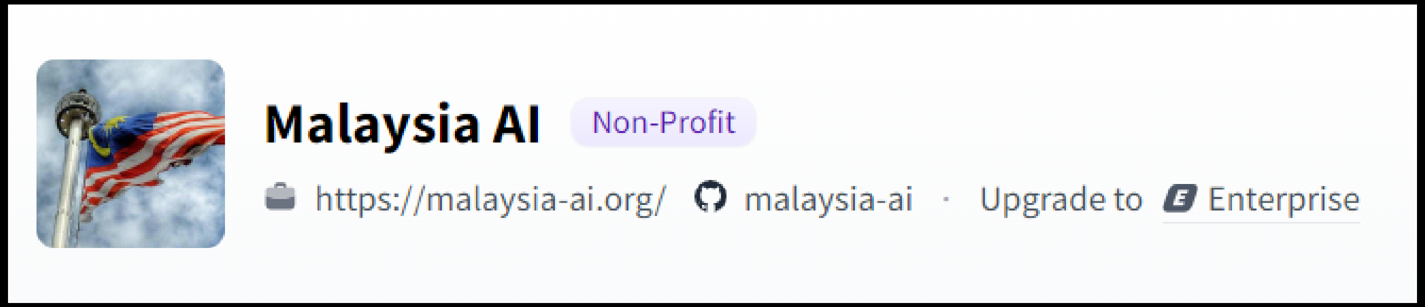
Especially to Mesolitica and Malaysia AI

for open-sourcing their models. In the same spirit, we've done the same in open-sourcing our datasets, and models for greater scrutiny and use as well.



Mesolitica Company

https://mesolitica.com/ X mesolitica mesolitica



Malaysia AI Non-Profit

https://malaysia-ai.org/ malaysia-ai · Upgrade to Enterprise