OPENAI WHISPER AND IT'S AMAZING POWER TO DO FINE-TUNING DEMONSTRATED ON MY MOTHER-TONGUE

Kurian Benoy

Sunday, October 1, 2023

OUTLINE

- What is OpenAI Whisper?
- Features of OpenAI Whisper
- What is Fine-tuning and how to fine-tune Whisper?
- About my mother tongue
- Methodology of benchmarking whisper models
- Results on benchmarking Whisper model
- Future Ideas & Conclusion

$WHOAMI

- AI Engineer & Team Lead @ sentient.io
- Volunteer @ Swathanthra Malayalam Computing(SMC)
- FOSS enthusiast
- Not affiliated to OpenAI

DISCLAIMER

- This talk is not generated.
- If I use something generated I will explicitly mark as from an LLM.

OPENAI WHISPER

- I think Whisper[1] is the most `under-rated model` released by OpenAI.
- It was open-sourced on September 21, 2022 by releasing the inference code and pre-trained model weights.

ABOUT OPENAI WHISPER MODEL

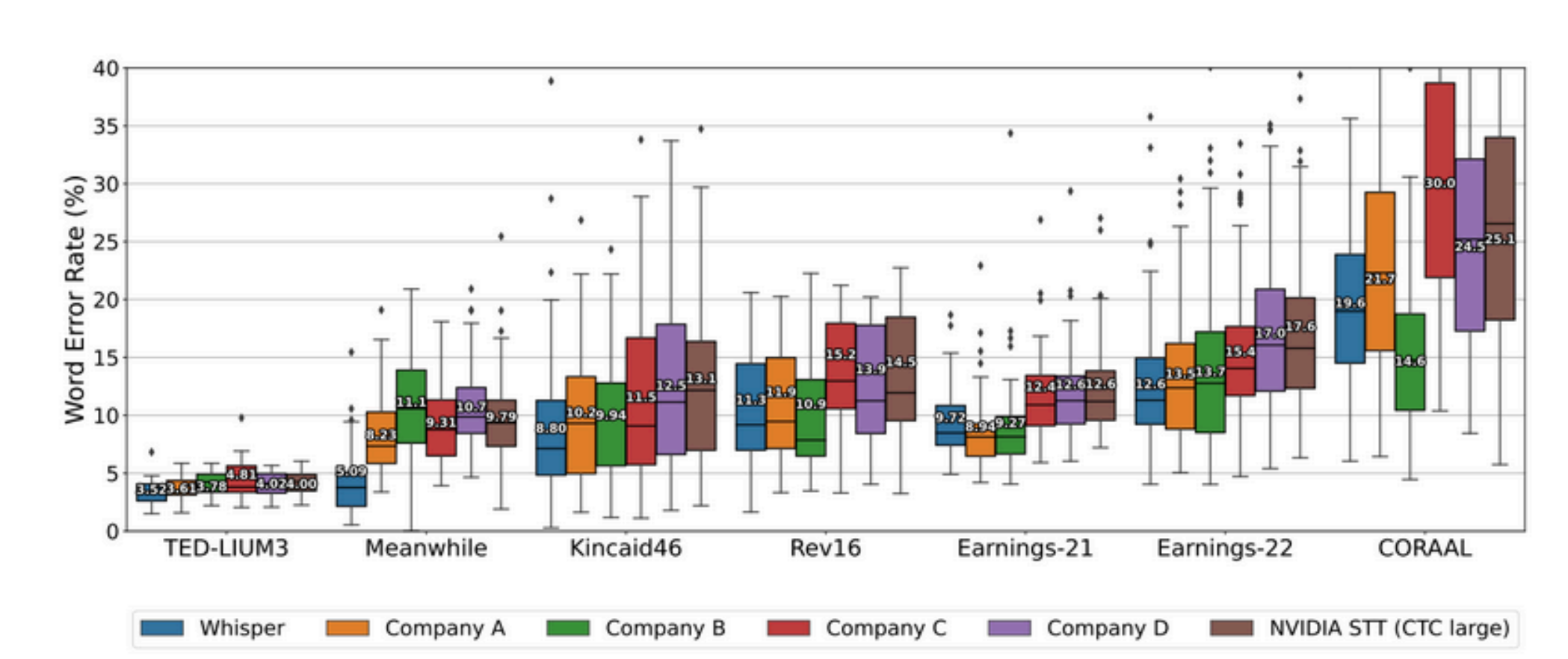
- Whisper is a computer program which can listen to people talking and write down what they say. (Automatic Speech Recognition Model)
- Whisper can understand people speaking different languages and can even translate what they say into English. (Supports transcription and translation to English)

Note: This was generated with GPT-4 with prompt: explain what is openai whisper to a

WHISPER MODELS

Size	Parameters	Required VRAM	Relative speed
tiny	39 M	~1 GB	~32x
base	74 M	~1 GB	~16x
small	244 M	~2 GB	~6x
medium	769 M	~5 GB	~2x
large	1550 M	~10 GB	1x

ENGLISH SPEECH RECOGNITION



Whisper is competitive with state of art commercial and open source systems.
Diagram from whisper research paper p.9

MULTI-LINGUAL SPEECH RECOGNITION

- Whisper model is trained on 99 languages
- OpenAI Whisper API supports just 57 languages as some languages performance are not really good.

RUNS IN ALMOST ANY DEVICE

- Since Whisper followed the open source route, whisper.cpp developed by Georgi Gerganov which is a port of OpenAI's Whisper model in C/C++.

- It supports the below platforms:

1. Mac OS (Intel and ARM)
2. iOS
3. Android
4. Linux/Free BSD
5. Web Assembly etc.

RUNS FASTER NOW

- JAX implementation of OpenAI Whisper model upto 70x speedup on TPU using Whisper JAX
- 4x faster with same accuracy using faster-whisper
- Useful transformer makes 2x faster than faster-whisper.

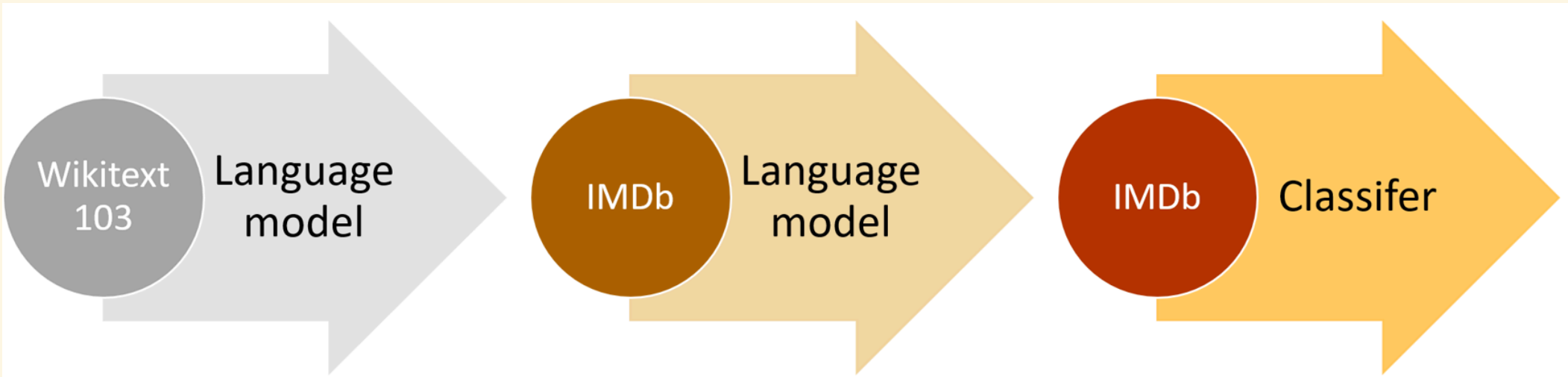
AWESOME COMMUNITY PLUGINS

- Word-level time stamps with whisper-timestamped,whisperX etc.
- Fine-Tune Whisper is achieving SOTA in lot of languages
- Speaker diarization
- Audio classification using OpenAI's Whisper

For more checkout awesome list by Sindre Sorhus

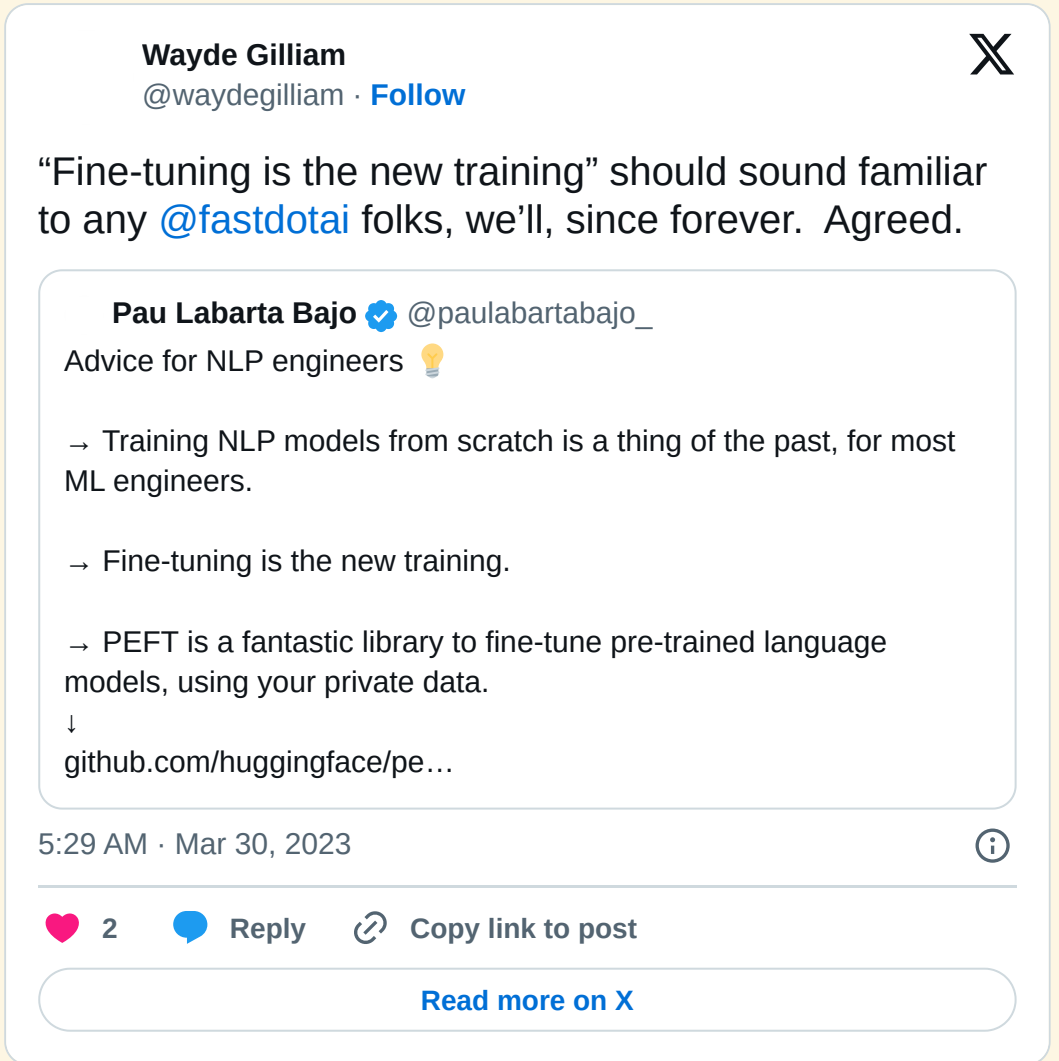
WHAT IS FINE TUNING?

Given a pre-trained model, which is a large model which is trained on a very specific task. If we want to fit it into our specific dataset we will train and use the pre-trained model to build a new model which works very well for our task.



Picture from fast.lesson covering steps in finetuning a text classifier model

FINE TUNING IS STILL RELEVANT



Wayde Gilliam
@waydegilliam · **Follow**

"Fine-tuning is the new training" should sound familiar to any @fastdotai folks, we'll, since forever. Agreed.

> **Pau Labarta Bajo** @paulabartabajo_
> Advice for NLP engineers 💡
>
> → Training NLP models from scratch is a thing of the past, for most ML engineers.
>
> → Fine-tuning is the new training.
>
> → PEFT is a fantastic library to fine-tune pre-trained language models, using your private data.
> ↓
> github.com/huggingface/pe…

5:29 AM · Mar 30, 2023

2 Reply Copy link to post

Read more on X

WHAT ARE STEPS FOR FINE-TUNING WHISPER?

Fine-Tune Whisper For Multilingual ASR with 🤗 Transformers

Published November 3, 2022

Update on GitHub

sanchit-gandhi
Sanchit Gandhi

Open in Colab

Fine-Tune Whisper For Multilingual ASR with 🤗 Transformers

WHAT ARE STEPS FOR FINE-TUNING WHISPER?

1. Preparing Environment
2. Load dataset
3. Prepare Feature Extractor, Tokenizer and Data
4. Training and evaluation
5. Building a demo(optional)

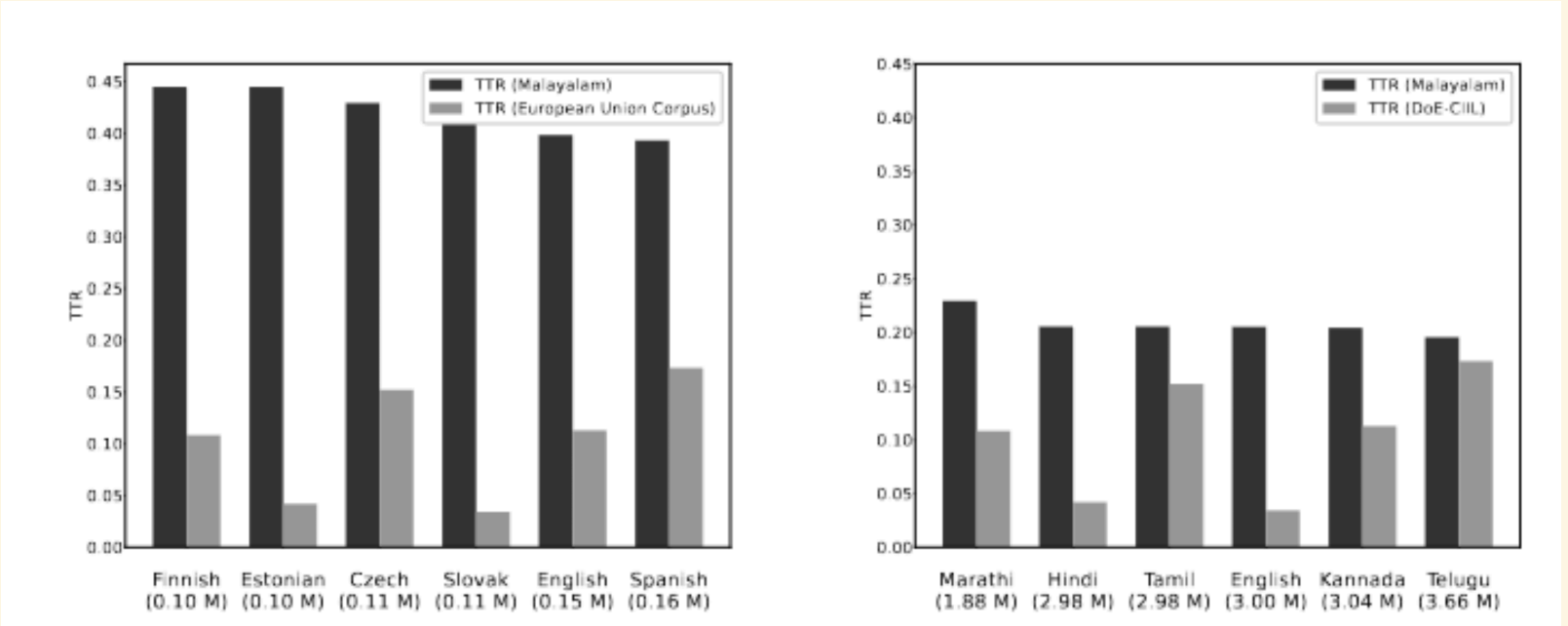
MALAYALAM PERFORMANCE IN WHISPER PAPER

Model	WER
tiny	102.7
base	122.9
small	104.8
medium	137.8
large-v1	107.1
large-v2	103.2

ABOUT MALAYALAM

- Malayalam is my mother tongue.
- Native speakers: 38+ million.(according to 2011 census)
- Spoken in: Kerala, Lakshadweep, Puducherry, wherever Malayalees are living.

MALAYALAM IS MORPHOLOGICALLY COMPLEX LANGUAGE



Comparison of Malayalam TTR with that of European Union Constitution Corpus and DoE-CIIL Corpus from K. Manohar et al.

Picture from Quantitative Analysis of the Morphological Complexity of Malayalam

WHISPER EVENT

- HuggingFace Team conducted a whisper fine tuning event for 2 weeks from 5th December 2022 to 19th December 2022. The results were out on 23rd December 2022.
- The goal was to to fine-tune the Whisper model to build state-of-the-art speech recognition systems in the languages of our choice 🗣

MALAYALAM MODELS PRODUCED IN WHISPER EVENT

- For the language Malayalam, the results are as follows:

Dataset

mozilla-foundation/common_voice_11_0

Language

ml: Malayalam

Split

test

Sorting Metric

wer
cer

🤗 Whisper Event: Final Leaderboard

This is the leaderboard for Common Voice 11 Malayalam (ml).

Please click on the model's name to be redirected to its model card.

Want to beat the leaderboard? Don't see your model here? Ensure...

model_id	wer	cer
thennal/whisper-medium-ml	11.49	-
anuragshas/whisper-large-v2-ml	25.48	-
parambharat/whisper-small-ml	25.80	-
DrishtiSharma/whisper-large-v2-malayalam	27.46	-
DrishtiSharma/whisper-large-v2-ml-700-steps	28.29	-
parambharat/whisper-base-ml	34.16	-
kurianbenoy/whisper_malayalam_largev2	41.70	-
parambharat/whisper-tiny-ml	45.72	-
5p33ch3xpr/Whisper-fineTuning-malayalam	68.74	-
kavaymanohar/whisper-small-malayalam	84.37	-

Malayalam models performance in whisper event according to leaderboard

WINNING MODELS IN MALAYALAM IN WHISPER EVENT

- The winning model for Common voice: `thennal/whisper-medium-ml`
- The winning model for Fleurs: `parambharath/whisper-small-ml`

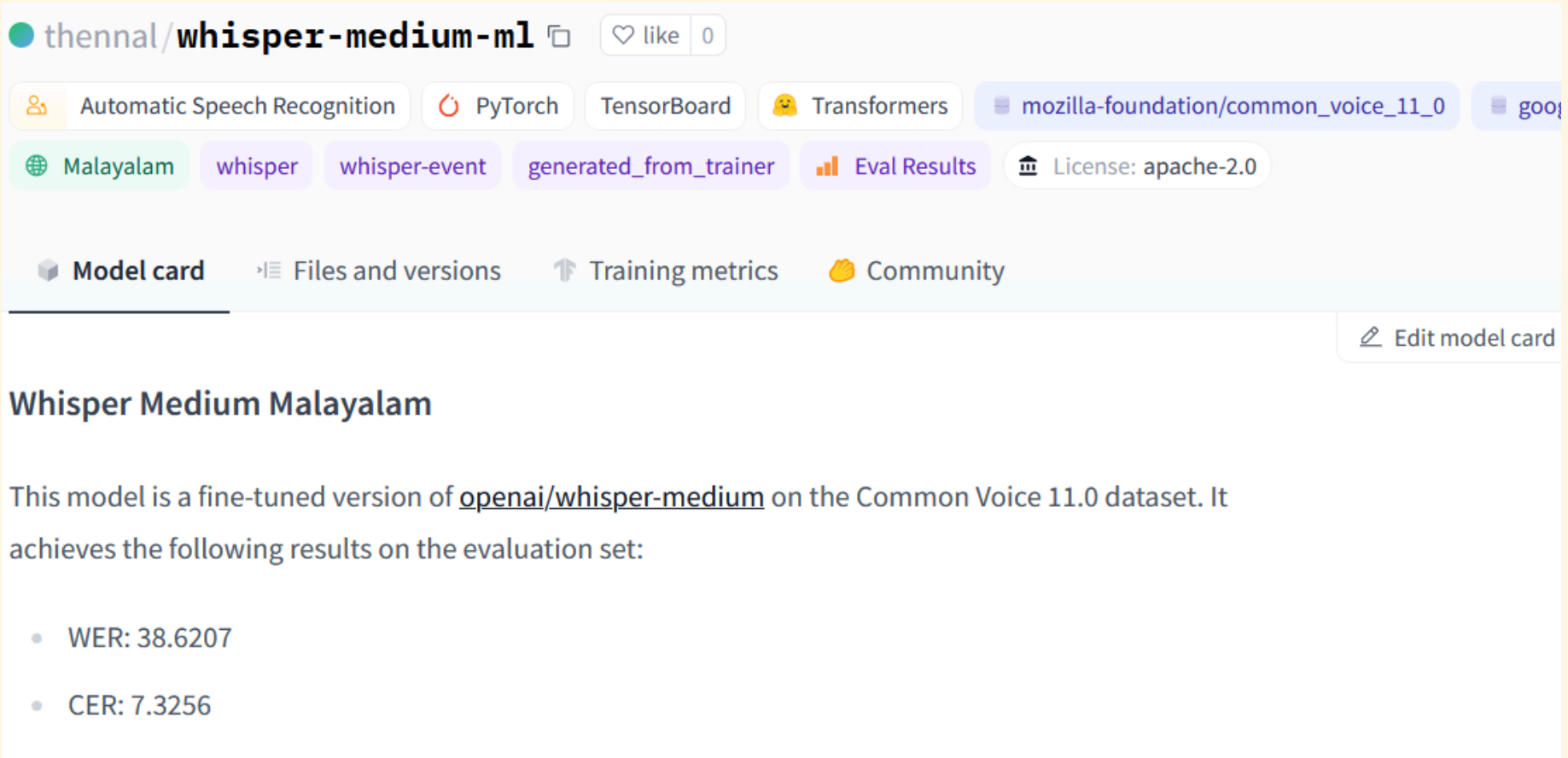
I WAS NOT CONVINCED

I was sceptical about the winning models becuase of:

1. Achieving 11% WER in Malayalam is astonishing.
2. In Malayalam there is not even a single yard stick to compare. Most of previous works were done in proprietary datasets and not open-sourced.
3. Malyalam is a morpohologically complex language. So even achieving 30% WER is a big deal.

I WAS NOT CONVINCED

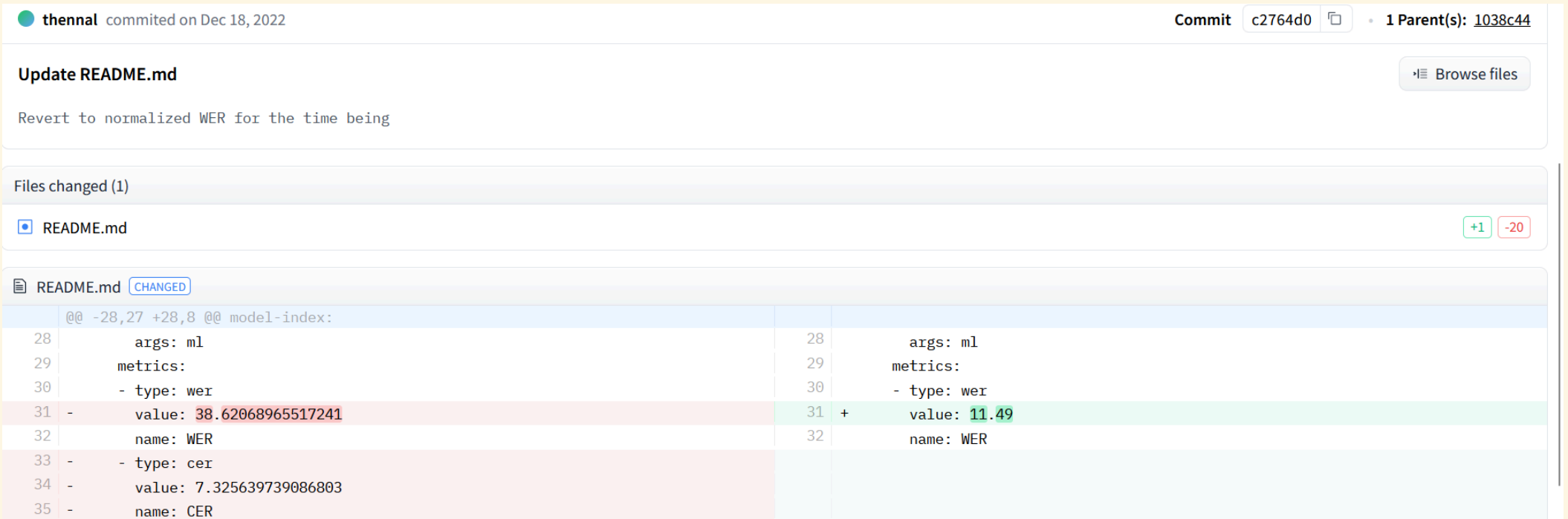
4. Didn't trust the Hugging Face way of evaluating models.



thennal/whisper-medium-ml model card readme

I WAS NOT CONVINCED

4. Didn't trust the Hugging Face way of evaluating models.



Last commit in thennal/whisper-medium-ml

METRICS FOR EVALUATING ASR MODELS

- ASR evaulation relies on comparission between ground-truth and ASR output.
- Common metrics for ASR evaluation which are popular and good enough[1] are :

1. Word Error Rate(WER)
2. Character Error Rate(CER)

To learn more about ASR evaluation check this blogpost by AWS

I WANTED TO BUILD SOMETHING NEW

- New github project for Malayalam ASR Benchmarking

Time for a new adventure

OBJECTIVE OF MY BENCHMARKING

- To test whether 10% WER was possible in available academic datasets.

Datasets

- Common Voice 11 malayalam subset
- SMC Malayalam Speech Corpus

METHODOLOGY FOR BENCHMARKING

1. Create as a python library so further whisper-based transformer models can be benchmark.
2. Calculate WER, CER, model size and time taken to benchmark the model for the listed datasets.
3. Build a reproducible approach, so results of benchmarking is stored as dataset.

LIBRARIES I USED

- Dependencies:
  - transformers
  - datasets
  - jiwer
  - whisper_normalizer
  - pandas
  - numerize
  - `librosa soundfile`

LIBRARIES I USED

- Development library:
  - nbdev
  - black
  - Jupyter Lab

LOADING THE DATASET FOR BENCHMARKING

```
def load_common_voice_malayalam_dataset():
    dataset = load_dataset(
            "mozilla-foundation/common_voice_11_0",
            "ml",
            split="test"
    )
    dataset = dataset.cast_column("audio", Audio(sampling_rate=1600
    dataset = dataset.map(normalise)
    dataset = dataset.filter(is_target_text_in_range, input_columns
    return dataset
```

BENCHMARKING A PARTICULAR MODEL WEIGHT IN COMMON VOICE

```
evaluate_whisper_model_common_voice(
    "openai/whisper-large",
    [], [], [], []
)
```

EVALUATING BENCHMARKING CODE

```
def evaluate_whisper_model_common_voice(
        model_name: str, # The model name
        werlist: List[float], # WER List
        cerlist: List[float],# CER list
        modelsizelist: List[str], # model size list
        timelist: List[float], # time(s) list
        bs:int =16, # batch size. Default value is 16.
)->None:
    whisper_asr = pipeline(
            "automatic-speech-recognition", model=model_name, devic
        )
    dataset = load_common_voice_malayalam_dataset()

    predictions = []
    references = []
    start = time.time()
    for out in whisper_asr(data(dataset), batch_size=bs):
        predictions.append(normalizer((out["text"])))
        references.append(normalizer(out["reference"][0]))
```

CALCULATING WER, CER

```
    ...
    df = pd.DataFrame({"predictions": predictions, "ground_truth":
    df["model_name"] = model_name
    df["wer"] = df.apply(lambda row: wer(normalizer(row["ground_tru
    df["cer"] = df.apply(lambda row: cer(normalizer(row["ground_tru
    df["total_time"] = end-start
    rwer = wer(references, predictions)
    rwer = round(100 * rwer, 2)
    werlist.append(rwer)
    print(f"The WER of model: {rwer}")

    rcer = cer(references, predictions)
    rcer = round(100 * rcer, 2)
    cerlist.append(rcer)
    print(f"The CER of model: {rcer}")
```

CALCULATING MODEL_SIZE AND STORING AS A DATASET

```
    ...
    print(f"The model size is: {get_model_size(whisper_asr.model)}"
    modelsizelist.append(get_model_size(whisper_asr.model))
    df["model_size"] = get_model_size(whisper_asr.model)

    save_name = model_name.split("/")
    print(save_name)
    df.to_parquet(f"{save_name[0]}_{save_name[1]}_commonvoice.parqu

    clear_gpu_memory()
```

BENCHMARKED MODELS

- Started with 6 fine-tuned models in Malayalam and compared it with 6 model versions released by OpenAI.

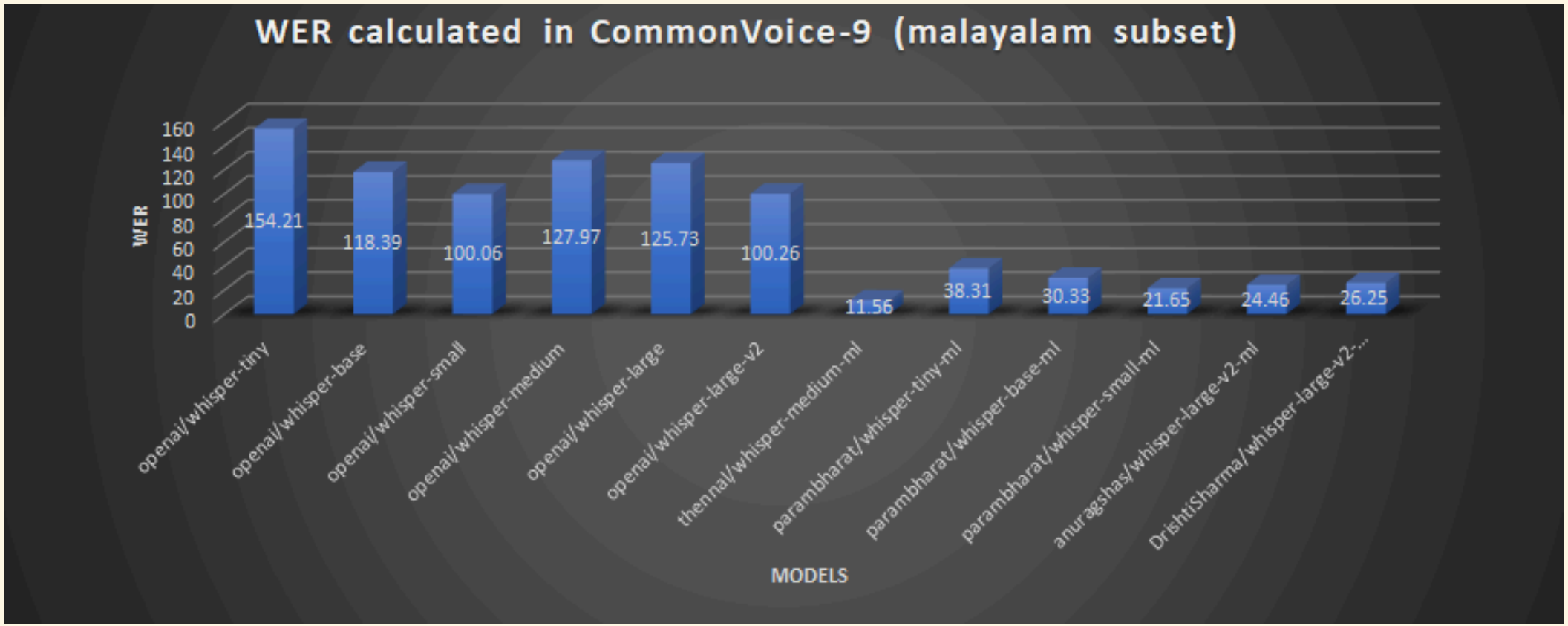
1. thennal/whisper-medium-ml
2. parambharat/whisper-tiny-ml
3. parambharat/whisper-base-ml
4. parambharat/whisper-small-ml
5. anuragshas/whisper-large-v2-ml
6. DrishtiSharma/whisper-large-v2-malayalam

RESULTS ON BENECHMARKING IN COMMON VOICE DATASET

MODEL NAME	WER	CER	MODEL SIZE	TIME(S)
openai/whisper-tiny	154.21	180.45	37.76M	22.277158
openai/whisper-base	118.39	131.08	72.59M	22.352587
openai/whisper-small	100.06	95.04	241.73M	25.442846
openai/whisper-medium	127.97	136.43	763.86M	53.880491
openai/whisper-large	125.73	139.62	1.54B	82.74608
openai/whisper-large-v2	100.26	93.6	1.54B	71.14292622
thennal/whisper-medium-ml	11.56	5.41	763.86M	924.979711
parambharat/whisper-tiny-ml	38.31	21.93	37.76M	59.535259
parambharat/whisper-base-ml	30.33	16.16	72.59M	96.419609
parambharat/whisper-small-ml	21.65	11.78	241.73M	273.555688
anuragshas/whisper-large-v2-ml	24.46	11.64	1.54B	1779.561592
DrishtiSharma/whisper-large-v2-malayalam	26.25	13.17	1.54B	1773.661774

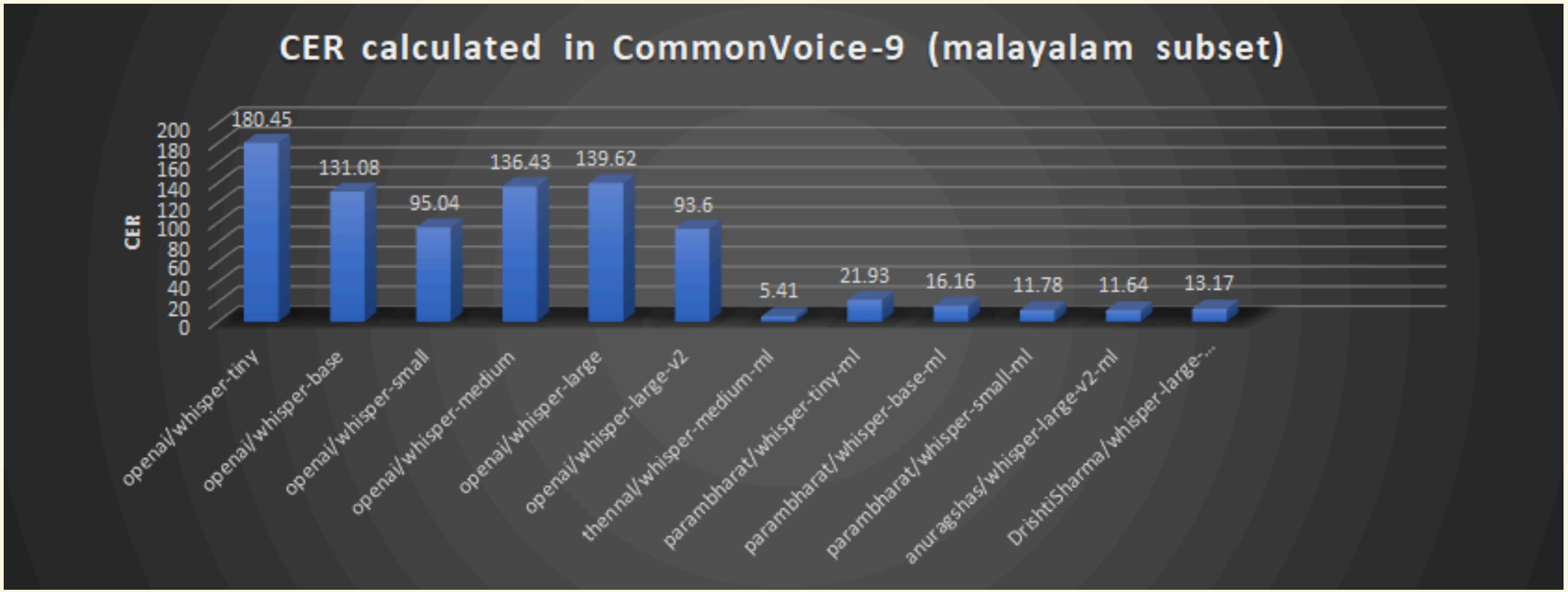
Output from benchmarking tool

WER IN COMMON VOICE DATASET

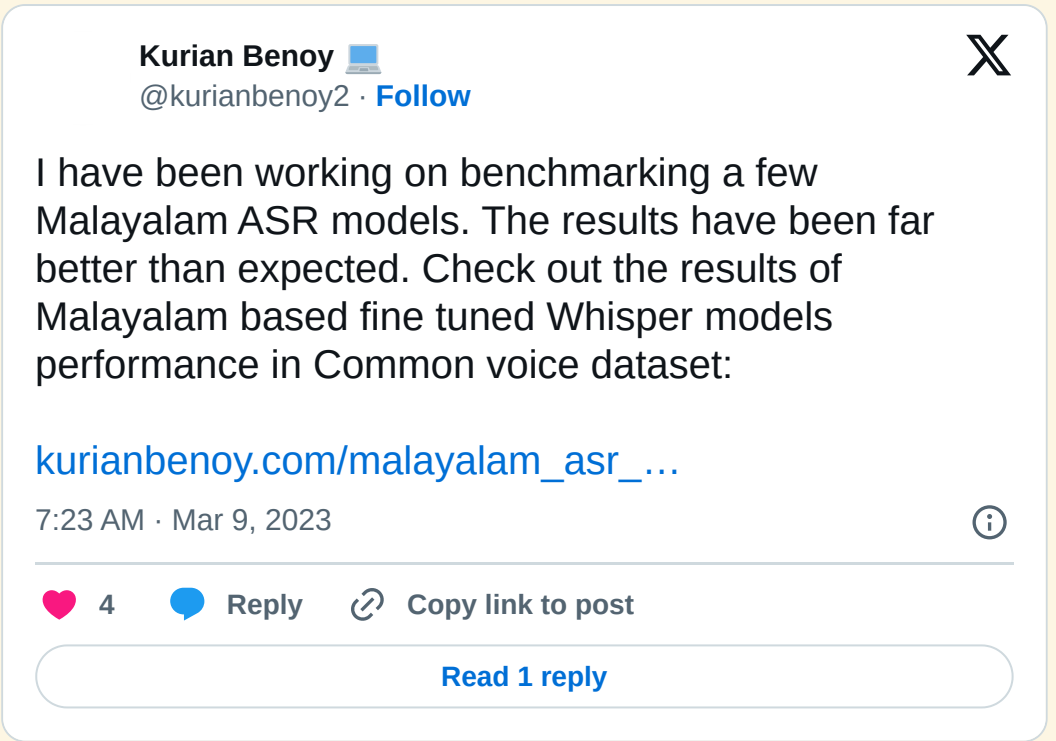


Word Error Rate in Common Voice-9 test split

CER IN COMMON VOICE DATASET



Character Error Rate in Common Voice-9 test split



Kurian Benoy 💻
@kurianbenoy2 · **Follow**

I have been working on benchmarking a few Malayalam ASR models. The results have been far better than expected. Check out the results of Malayalam based fine tuned Whisper models performance in Common voice dataset:

kurianbenoy.com/malayalam_asr_…

7:23 AM · Mar 9, 2023

4 Reply Copy link to post

Read 1 reply

Kavya Manohar (കാവ്യ)
@kavya_manohar · **Follow**

There has never been a benchmark for comparing Malayalam ASR models. Thank you @kurianbenoy2

Kurian Benoy 💻 @kurianbenoy2

Replying to @kurianbenoy2

I am building a benchmarking tool to benchmark few more datasets. Some popular open source datasets like @smcproject Malayalam Speech Corpus will be benchmarked soon.

You can find code here:
github.com/kurianbenoy/ma…

2:33 PM · Mar 9, 2023

3 Reply Copy link to post

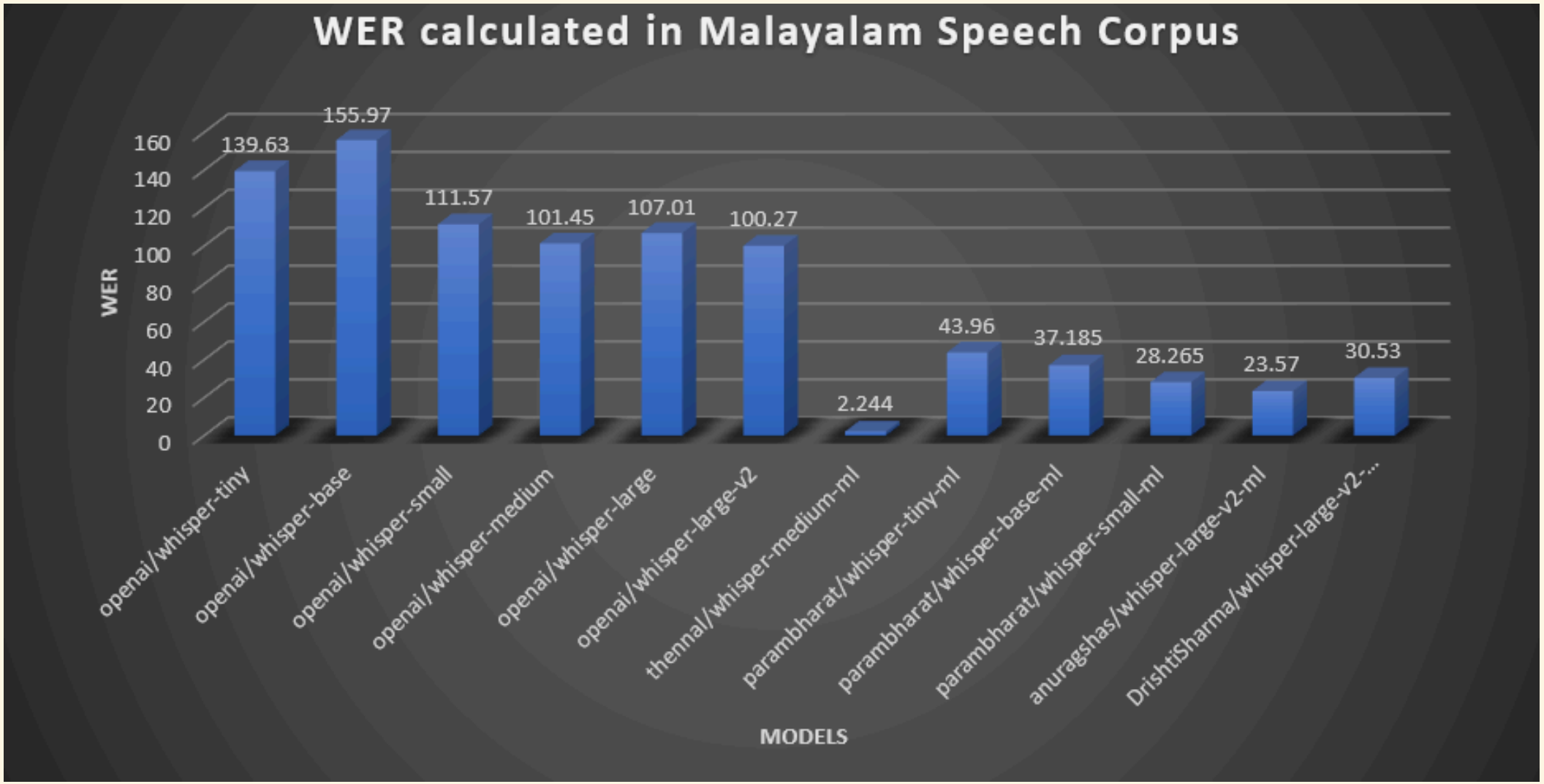
Read more on X

RESULTS ON BENECHMARKING IN MALAYALAM SPEECH CORPUS DATASET

MODEL NAME	WER	CER	MODEL SIZE	TIME(seconds)
openai/whisper-tiny	139.63	177.3	37.76M	375.532
openai/whisper-base	155.97	200.05	72.59M	448.95
openai/whisper-small	111.57	123.7	241.73M	479.736
openai/whisper-medium	101.45	104.23	763.86M	672.291
openai/whisper-large	107.01	113.62	1.54B	1067.557
openai/whisper-large-v2	100.27	102.4	1.54B	1040.25
thennal/whisper-medium-ml	2.244	1.247	763.86M	8736.731
parambharat/whisper-tiny-ml	43.96	25.78	37.76M	727.57
parambharat/whisper-base-ml	37.185	21.389	72.59M	1124.314
parambharat/whisper-small-ml	28.265	15.379	241.73M	2893.445
anuragshas/whisper-large-v2-ml	23.57	12.33	1.54B	10467.876
DrishtiSharma/whisper-large-v2-malayalam	30.53	19.81	1.54B	10067.01

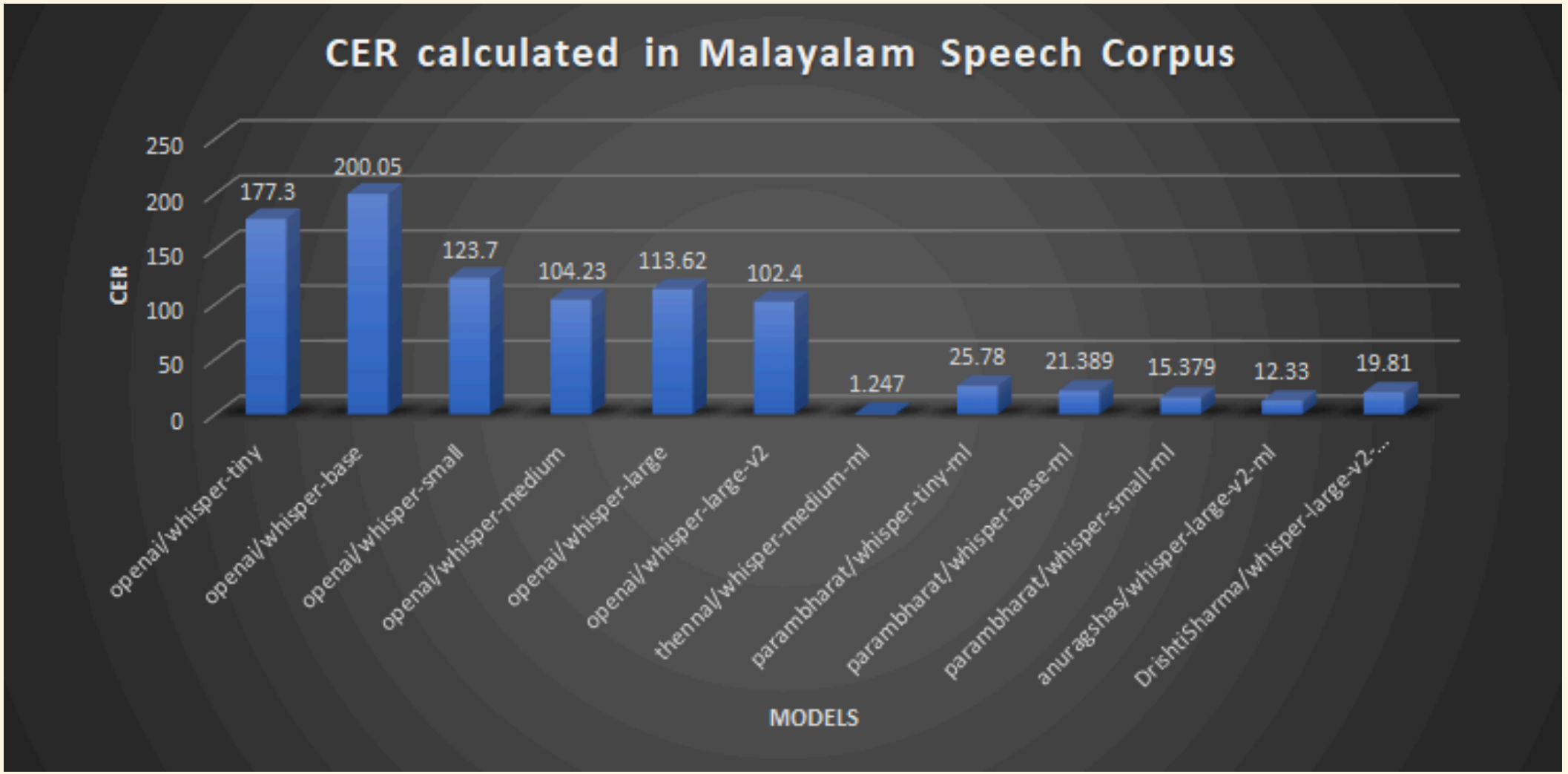
Output from benchmarking tool

WER IN MALAYALAM SPEECH CORPUS



Word Error Rate in MSC

CER IN MALAYALAM SPEECH CORPUS



Character Error rate in MSC

LINKS TO PROJECT

Github project

https://github.com/kurianbenoy/malayalam_asr_benchmarking

LINKS TO PROJECT

Benchmarking results

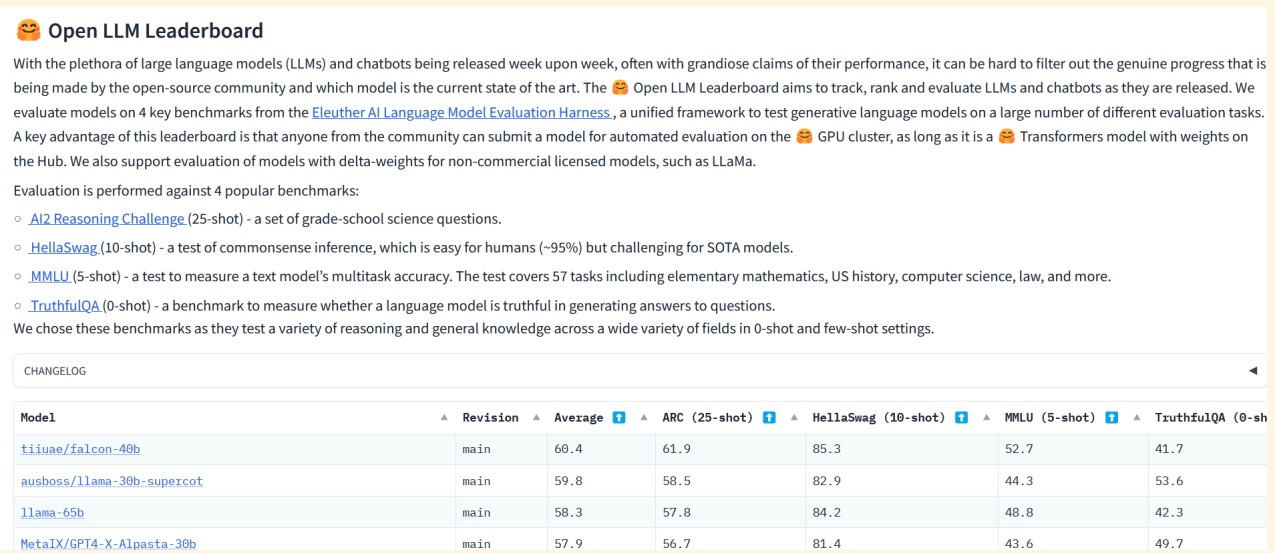
- Results on SMC Malayalam Speech corpus

  https://huggingface.co/datasets/kurianbenoy/malayalam_msc_benchmarking
- Results on Common Voice 11

  https://huggingface.co/datasets/kurianbenoy/malayalam_common_voice_benchmarking

FUTURE IDEAS FOR BENCHMARKING

- Something very similar to OpenLLM Leaderboard with results of latest malayalam speech models.
- Should include results for ASR models based on other architectures like Kaldi, Meta's MMS, Wav2Vec etc.



🤗 Open LLM Leaderboard

With the plethora of large language models (LLMs) and chatbots being released week upon week, often with grandiose claims of their performance, it can be hard to filter out the genuine progress that is being made by the open-source community and which model is the current state of the art. The 🤗 Open LLM Leaderboard aims to track, rank and evaluate LLMs and chatbots as they are released. We evaluate models on 4 key benchmarks from the Eleuther AI Language Model Evaluation Harness , a unified framework to test generative language models on a large number of different evaluation tasks. A key advantage of this leaderboard is that anyone from the community can submit a model for automated evaluation on the 🤗 GPU cluster, as long as it is a 🤗 Transformers model with weights on the Hub. We also support evaluation of models with delta-weights for non-commercial licensed models, such as LLaMa.

Evaluation is performed against 4 popular benchmarks:

- AI2 Reasoning Challenge (25-shot) - a set of grade-school science questions.
- HellaSwag (10-shot) - a test of commonsense inference, which is easy for humans (~95%) but challenging for SOTA models.
- MMLU (5-shot) - a test to measure a text model's multitask accuracy. The test covers 57 tasks including elementary mathematics, US history, computer science, law, and more.
- TruthfulQA (0-shot) - a benchmark to measure whether a language model is truthful in generating answers to questions.

We chose these benchmarks as they test a variety of reasoning and general knowledge across a wide variety of fields in 0-shot and few-shot settings.

CHANGELOG

Model	Revision	Average	ARC (25-shot)	HellaSwag (10-shot)	MMLU (5-shot)	TruthfulQA (0-sh
tiiuae/falcon-40b	main	60.4	61.9	85.3	52.7	41.7
ausboss/llama-30b-supercot	main	59.8	58.5	82.9	44.3	53.6
llama-65b	main	58.3	57.8	84.2	48.8	42.3
MetaIX/GPT4-X-Alpasta-30b	main	57.9	56.7	81.4	43.6	49.7

Open LLM leaderboard in huggingface spaces

CONCLUSION

- In Malayalam we have achieved phenomenal results for fine tuned whisper models.
- The best model after benchmarking is: `thennal/whisper-medium-ml`
- You can also do it in your own language especially if it is a low resource language.

THANKS TO

1. OpenAI team - Alec Radford, Jong Wook Kim, Christine McLeavey etc. other authors of Whisper paper
2. Creators of CTranslate2 and faster-whisper - Guillaume Klein
3. HuggingFace team - Sanchit Gandhi, Nicolas Patry, Vaibhav Srivastav etc.
4. Kavya Manohar
5. Santhosh Thottingal
6. Thennal D K
7. AbdulMajedRaja RS
8. Georgi Gerganov
9. Ramsri Goutham
10. Wayde Gilliam
11. Other members in SMC.
12. Jarvis Labs

TRIBUTES

Remembering my first Python teacher