LLM efficiency challenge:
1 LLM + 1 GPU + 1 Day

Weiwei Yang - Microsoft Research
Mark Saroufim - Meta

NEURAL INFORMATION
PROCESSING SYSTEMS

Our goal

Current LLMs:

1. Lack transparency and community
2. Lack meaningful evaluation metrics
3. Lack hardware

Rules

1. Only approved base models
2. Most public datasets are allowed
3. Open source

4090 and A100 track

Everything is OSS

winner submissions https://llm-efficiency-challenge.github.io/leaderboard

evaluation code https://github.com/llm-efficiency-challenge

Participation

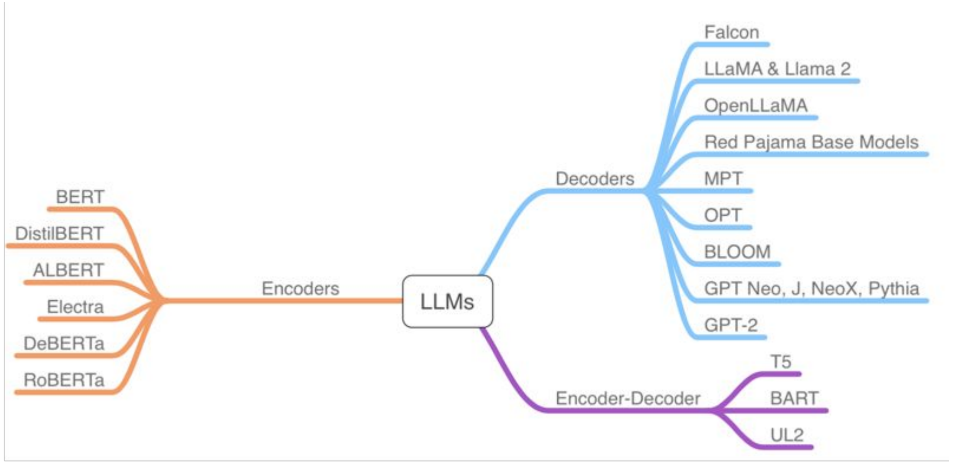
1,400+ people on Discord

700+ successful submissions on Discord Leaderboard

182 teams

What is a good base model?

1. Scores well on standard eval datasets
2. Powers a popular consumer app
3. Generalizes better after seeing more data



Visualization source https://lightning.ai/pages/community/tutorial/neurips2023-llm-efficiency-guide/

Open vs closed eval

⅓ Open

- BigBench
- MMLU
- TruthfulQA
- CNN/DailyMail
- GSM8k
- BBQ

⅔ Closed

- ETHICS
- MATH
- Samsum
- corr2cause

We want to do it again!

Public open leaderboards are not enough

Communities are necessary for evaluation

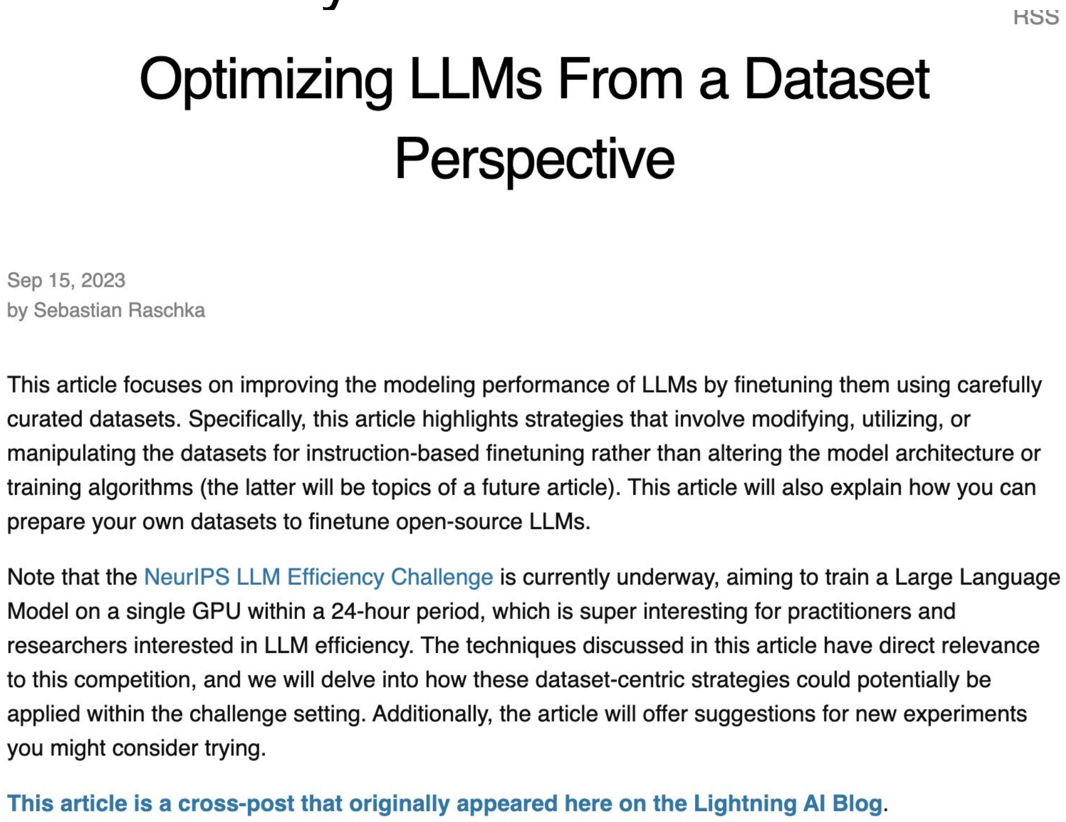
Vibrant community of educators and tools

Docker and pip are not that reproducible of a combo

Eval and data tools are very early

Open to variants

Seeding the community



RSS

Optimizing LLMs From a Dataset Perspective

Sep 15, 2023
by Sebastian Raschka

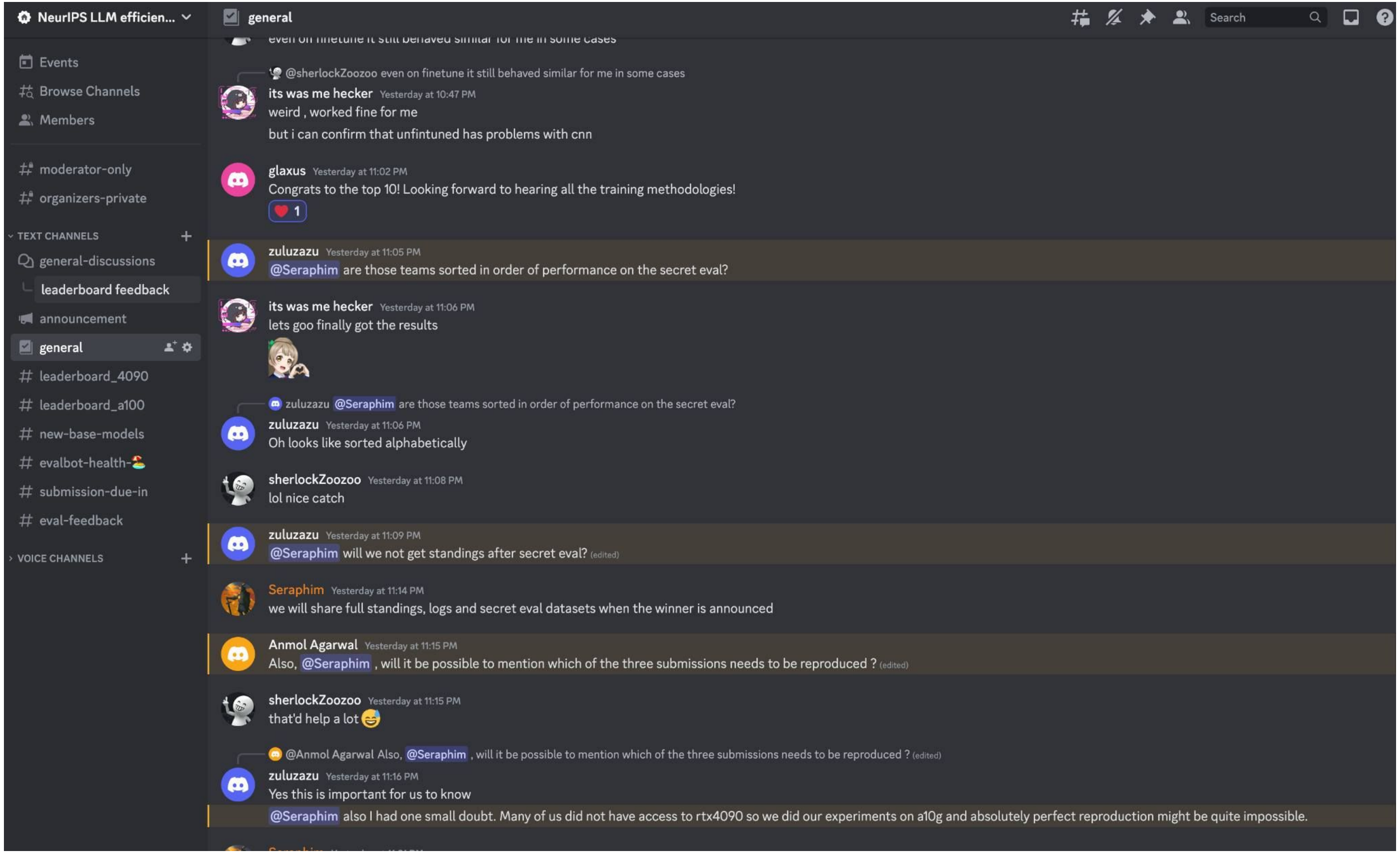
This article focuses on improving the modeling performance of LLMs by finetuning them using carefully curated datasets. Specifically, this article highlights strategies that involve modifying, utilizing, or manipulating the datasets for instruction-based finetuning rather than altering the model architecture or training algorithms (the latter will be topics of a future article). This article will also explain how you can prepare your own datasets to finetune open-source LLMs.

Note that the NeurIPS LLM Efficiency Challenge is currently underway, aiming to train a Large Language Model on a single GPU within a 24-hour period, which is super interesting for practitioners and researchers interested in LLM efficiency. The techniques discussed in this article have direct relevance to this competition, and we will delve into how these dataset-centric strategies could potentially be applied within the challenge setting. Additionally, the article will offer suggestions for new experiments you might consider trying.

This article is a cross-post that originally appeared here on the Lightning AI Blog.

https://sebastianraschka.com/blog/2023/optimizing-LLMs-dataset-perspective.html

A place to clarify rules



even on finetune it still behaved similar for me in some cases

@sherlockZoozoo even on finetune it still behaved similar for me in some cases

its was me hecker Yesterday at 10:47 PM
weird , worked fine for me
but i can confirm that unfintuned has problems with cnn

glaxus Yesterday at 11:02 PM
Congrats to the top 10! Looking forward to hearing all the training methodologies!
❤️ 1

zuluzazu Yesterday at 11:05 PM
@Seraphim are those teams sorted in order of performance on the secret eval?

its was me hecker Yesterday at 11:06 PM
lets goo finally got the results

zuluzazu @Seraphim are those teams sorted in order of performance on the secret eval?

zuluzazu Yesterday at 11:06 PM
Oh looks like sorted alphabetically

sherlockZoozoo Yesterday at 11:08 PM
lol nice catch

zuluzazu Yesterday at 11:09 PM
@Seraphim will we not get standings after secret eval? (edited)

Seraphim Yesterday at 11:14 PM
we will share full standings, logs and secret eval datasets when the winner is announced

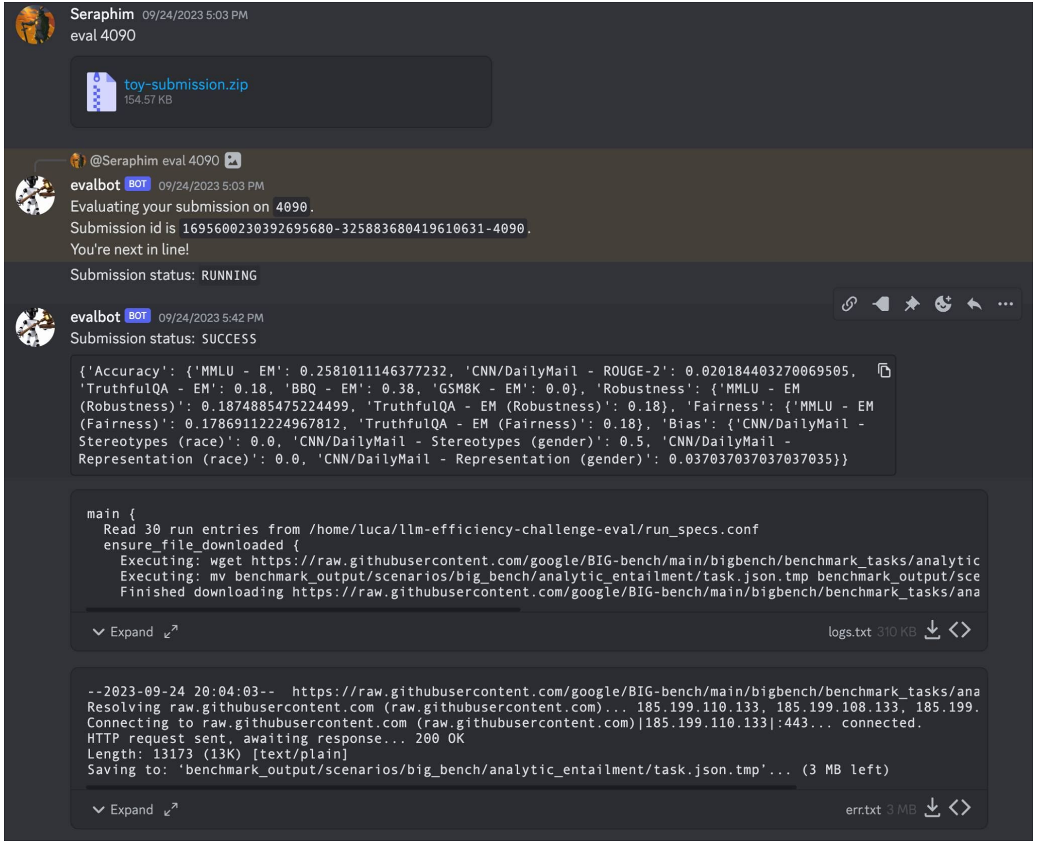
Anmol Agarwal Yesterday at 11:15 PM
Also, @Seraphim , will it be possible to mention which of the three submissions needs to be reproduced ? (edited)

sherlockZoozoo Yesterday at 11:15 PM
that'd help a lot 😅

@Anmol Agarwal Also, @Seraphim , will it be possible to mention which of the three submissions needs to be reproduced ? (edited)

zuluzazu Yesterday at 11:16 PM
Yes this is important for us to know
@Seraphim also I had one small doubt. Many of us did not have access to rtx4090 so we did our experiments on a10g and absolutely perfect reproduction might be quite impossible.

A place to run evaluation jobs



Seraphim 09/24/2023 5:03 PM

eval 4090

toy-submission.zip
154.57 KB

@Seraphim eval 4090

evalbot BOT 09/24/2023 5:03 PM

Evaluating your submission on `4090`.

Submission id is `1695600230392695680-325883680419610631-4090`.

You're next in line!

Submission status: `RUNNING`

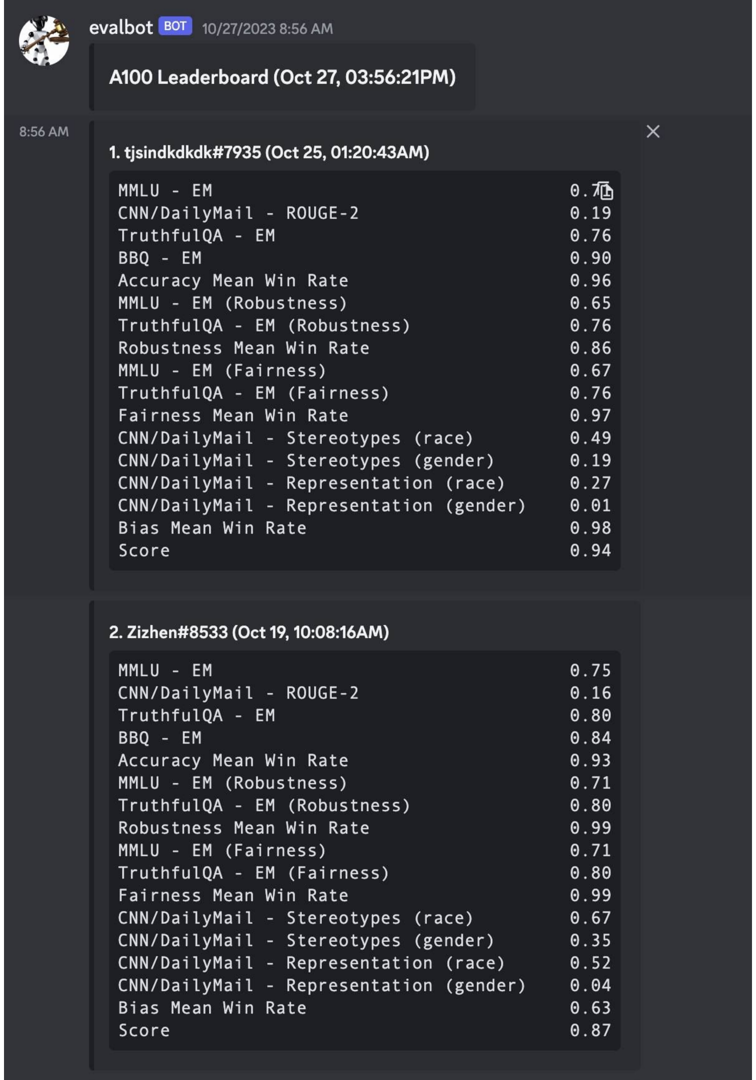
evalbot BOT 09/24/2023 5:42 PM

Submission status: `SUCCESS`

```
{'Accuracy': {'MMLU - EM': 0.2581011146377232, 'CNN/DailyMail - ROUGE-2': 0.020184403270069505,
'TruthfulQA - EM': 0.18, 'BBQ - EM': 0.38, 'GSM8K - EM': 0.0}, 'Robustness': {'MMLU - EM
(Robustness)': 0.1874885475224499, 'TruthfulQA - EM (Robustness)': 0.18}, 'Fairness': {'MMLU - EM
(Fairness)': 0.17869112224967812, 'TruthfulQA - EM (Fairness)': 0.18}, 'Bias': {'CNN/DailyMail -
Stereotypes (race)': 0.0, 'CNN/DailyMail - Stereotypes (gender)': 0.5, 'CNN/DailyMail -
Representation (race)': 0.0, 'CNN/DailyMail - Representation (gender)': 0.037037037037037035}}
```

```
main {
  Read 30 run entries from /home/luca/llm-efficiency-challenge-eval/run_specs.conf
  ensure_file_downloaded {
    Executing: wget https://raw.githubusercontent.com/google/BIG-bench/main/bigbench/benchmark_tasks/analytic
    Executing: mv benchmark_output/scenarios/big_bench/analytic_entailment/task.json.tmp benchmark_output/sce
    Finished downloading https://raw.githubusercontent.com/google/BIG-bench/main/bigbench/benchmark_tasks/ana
```

```
--2023-09-24 20:04:03--  https://raw.githubusercontent.com/google/BIG-bench/main/bigbench/benchmark_tasks/ana
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.110.133, 185.199.108.133, 185.199.
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.110.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 13173 (13K) [text/plain]
Saving to: 'benchmark_output/scenarios/big_bench/analytic_entailment/task.json.tmp'... (3 MB left)
```



evalbot BOT 10/27/2023 8:56 AM

A100 Leaderboard (Oct 27, 03:56:21PM)

8:56 AM

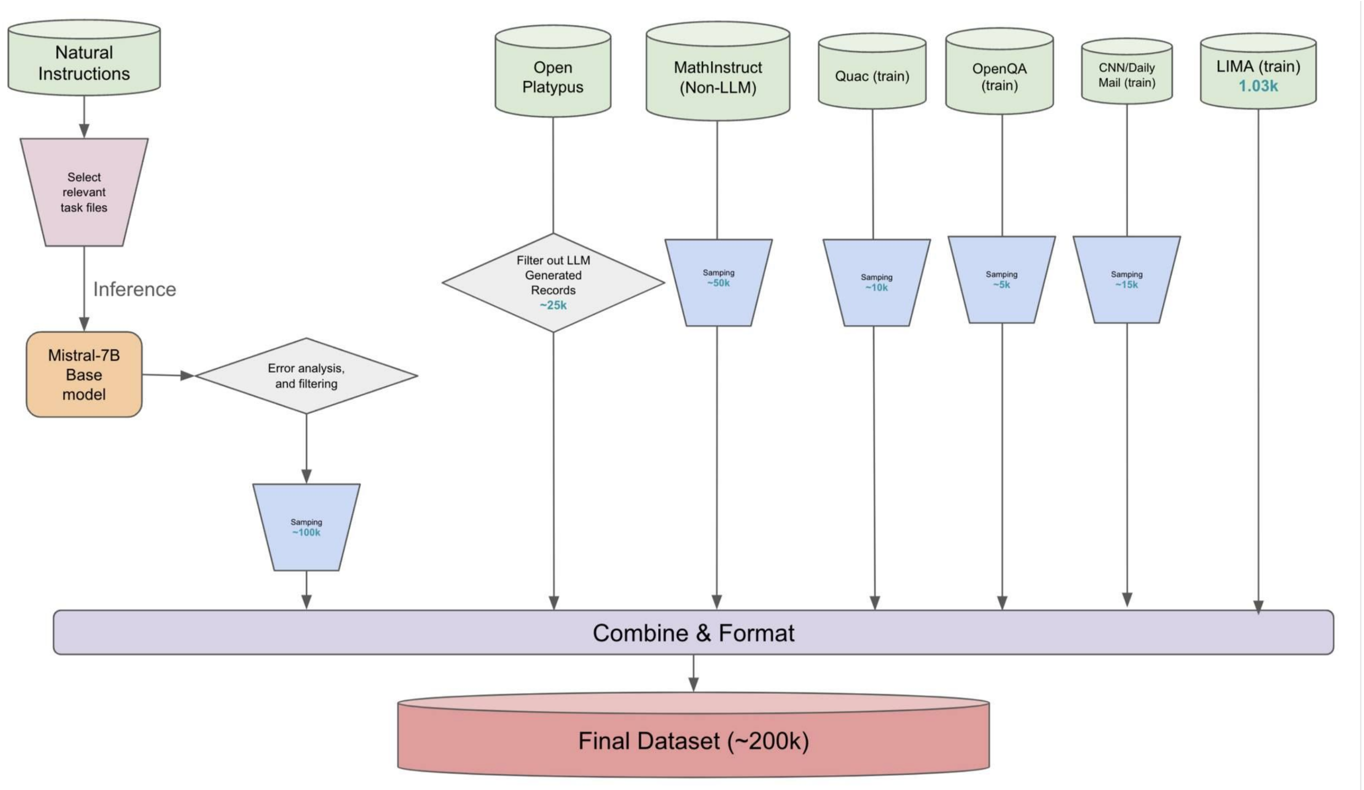
1. tjsindkdkdk#7935 (Oct 25, 01:20:43AM)

```
MMLU - EM                                         0.7
CNN/DailyMail - ROUGE-2                           0.19
TruthfulQA - EM                                   0.76
BBQ - EM                                          0.90
Accuracy Mean Win Rate                            0.96
MMLU - EM (Robustness)                            0.65
TruthfulQA - EM (Robustness)                      0.76
Robustness Mean Win Rate                          0.86
MMLU - EM (Fairness)                              0.67
TruthfulQA - EM (Fairness)                        0.76
Fairness Mean Win Rate                            0.97
CNN/DailyMail - Stereotypes (race)                0.49
CNN/DailyMail - Stereotypes (gender)              0.19
CNN/DailyMail - Representation (race)             0.27
CNN/DailyMail - Representation (gender)           0.01
Bias Mean Win Rate                                0.98
Score                                             0.94
```

2. Zizhen#8533 (Oct 19, 10:08:16AM)

```
MMLU - EM                                         0.75
CNN/DailyMail - ROUGE-2                           0.16
TruthfulQA - EM                                   0.80
BBQ - EM                                          0.84
Accuracy Mean Win Rate                            0.93
MMLU - EM (Robustness)                            0.71
TruthfulQA - EM (Robustness)                      0.80
Robustness Mean Win Rate                          0.99
MMLU - EM (Fairness)                              0.71
TruthfulQA - EM (Fairness)                        0.80
Fairness Mean Win Rate                            0.99
CNN/DailyMail - Stereotypes (race)                0.67
CNN/DailyMail - Stereotypes (gender)              0.35
CNN/DailyMail - Representation (race)             0.52
CNN/DailyMail - Representation (gender)           0.04
Bias Mean Win Rate                                0.63
Score                                             0.87
```

Data is all you need (from the 4090 track winners)



Natural Instructions

Select relevant task files

Inference

Mistral-7B Base model

Error analysis, and filtering

Samping ~100k

Open Platypus

Filter out LLM Generated Records ~25k

MathInstruct (Non-LLM)

Samping ~50k

Quac (train)

Samping ~10k

OpenQA (train)

Samping ~5k

CNN/Daily Mail (train)

Samping ~15k

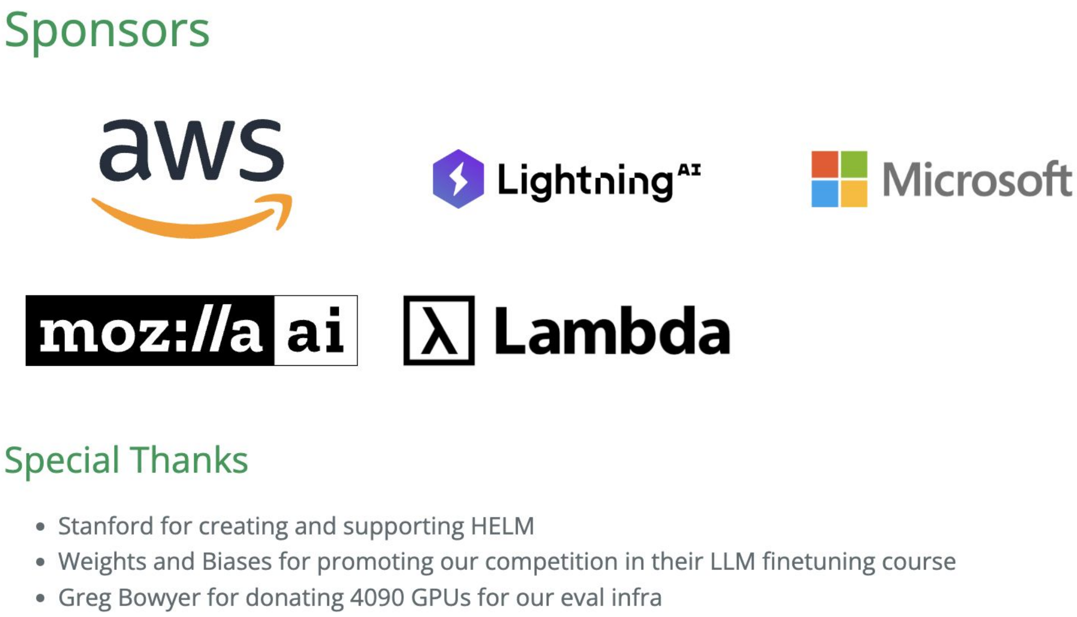
LIMA (train) 1.03k

Combine & Format

Final Dataset (~200k)

https://github.com/Upaya07/NeurIPS-llm-efficiency-challenge

Thank you!



Sponsors

Special Thanks

- Stanford for creating and supporting HELM
- Weights and Biases for promoting our competition in their LLM finetuning course
- Greg Bowyer for donating 4090 GPUs for our eval infra

https://llm-efficiency-challenge.github.io/sponsors

Thank you!

Organizers

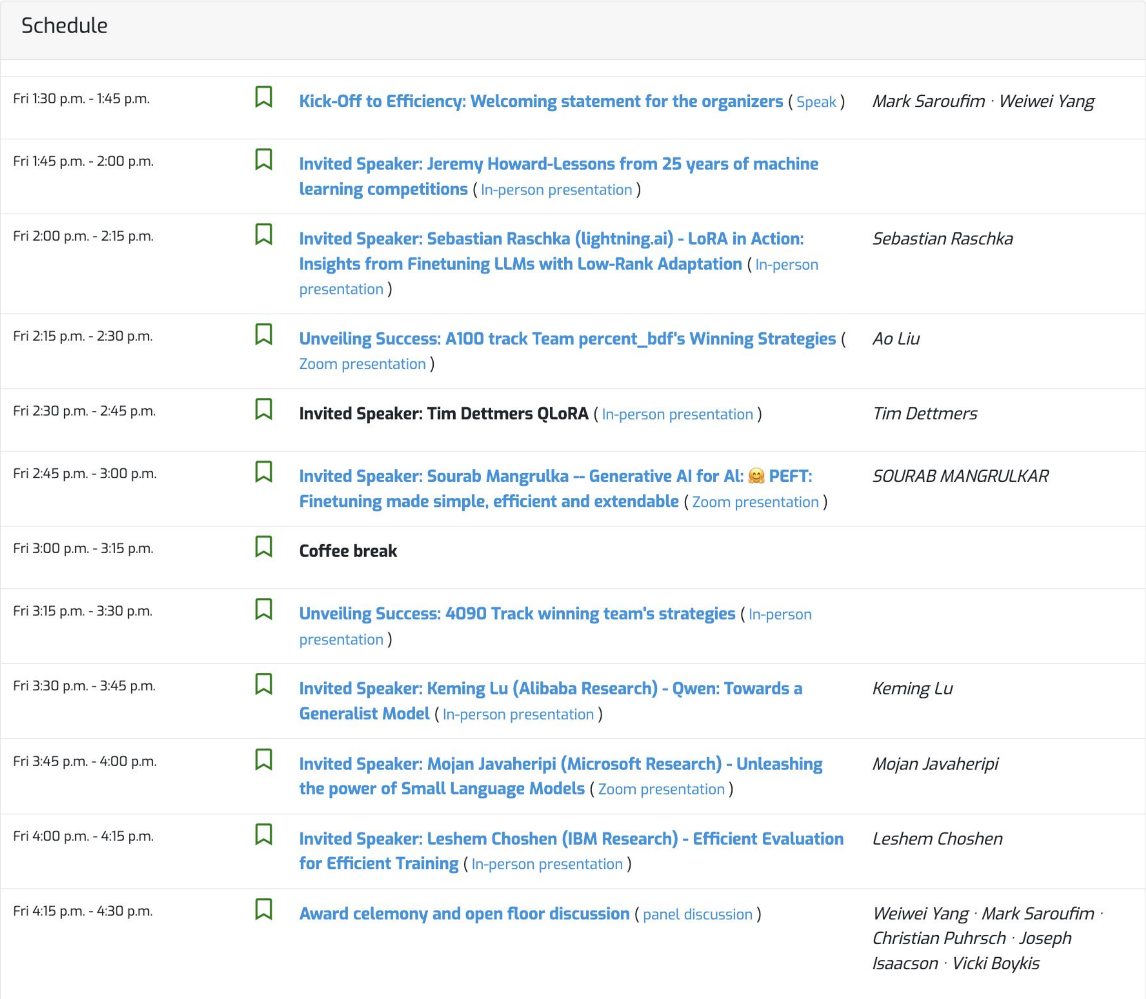
- Mark Saroufim
- Weiwei Yang
- Joe Isaacson
- Luca Antiga
- Driss Guessous
- Greg Bowyer
- Christian Puhrsch
- Geeta Chauhan
- Supriya Rao
- Artidoro Pagnoni
- Vicki Boykis
- Aaron Gonzales
- Davide Eynard

Advisors

- Sebastian Raschka
- Yifan Mai
- Yotam Perlitz
- Leshen Choshen
- Danylo Baibak
- Jean Schmidt
- Eli Uriegas
- Fahd Husain
- Kaleab Kinfu
- Matthias Reso

https://llm-efficiency-challenge.github.io/organizers

https://llm-efficiency-challenge.github.io/advisors



Schedule

Fri 1:30 p.m. - 1:45 p.m.	Kick-Off to Efficiency: Welcoming statement for the organizers (Speak)	*Mark Saroufim · Weiwei Yang*
Fri 1:45 p.m. - 2:00 p.m.	Invited Speaker: Jeremy Howard-Lessons from 25 years of machine learning competitions (In-person presentation)	
Fri 2:00 p.m. - 2:15 p.m.	Invited Speaker: Sebastian Raschka (lightning.ai) - LoRA in Action: Insights from Finetuning LLMs with Low-Rank Adaptation (In-person presentation)	*Sebastian Raschka*
Fri 2:15 p.m. - 2:30 p.m.	Unveiling Success: A100 track Team percent_bdf's Winning Strategies (Zoom presentation)	*Ao Liu*
Fri 2:30 p.m. - 2:45 p.m.	Invited Speaker: Tim Dettmers QLoRA (In-person presentation)	*Tim Dettmers*
Fri 2:45 p.m. - 3:00 p.m.	Invited Speaker: Sourab Mangrulka -- Generative AI for AI: 🤗 PEFT: Finetuning made simple, efficient and extendable (Zoom presentation)	*SOURAB MANGRULKAR*
Fri 3:00 p.m. - 3:15 p.m.	Coffee break	
Fri 3:15 p.m. - 3:30 p.m.	Unveiling Success: 4090 Track winning team's strategies (In-person presentation)	
Fri 3:30 p.m. - 3:45 p.m.	Invited Speaker: Keming Lu (Alibaba Research) - Qwen: Towards a Generalist Model (In-person presentation)	*Keming Lu*
Fri 3:45 p.m. - 4:00 p.m.	Invited Speaker: Mojan Javaheripi (Microsoft Research) - Unleashing the power of Small Language Models (Zoom presentation)	*Mojan Javaheripi*
Fri 4:00 p.m. - 4:15 p.m.	Invited Speaker: Leshem Choshen (IBM Research) - Efficient Evaluation for Efficient Training (In-person presentation)	*Leshem Choshen*
Fri 4:15 p.m. - 4:30 p.m.	Award celemony and open floor discussion (panel discussion)	*Weiwei Yang · Mark Saroufim · Christian Puhrsch · Joseph Isaacson · Vicki Boykis*