Lightning AI

Creators of PyTorch Lightning

The current scenario of LLM evaluation

Anindyadeep Sannigrahi
Data science engineering at PremAI-io

November 2023

Scenarios, Task Benchmark Dataset and Metrics

A Scenario is a specific context/setting or a condition under which the LLM’s performance is assessed and tested.

Example:

- Question Answering
- Reasoning
- Machine Translation
- Text Generation and Natural Language Understanding

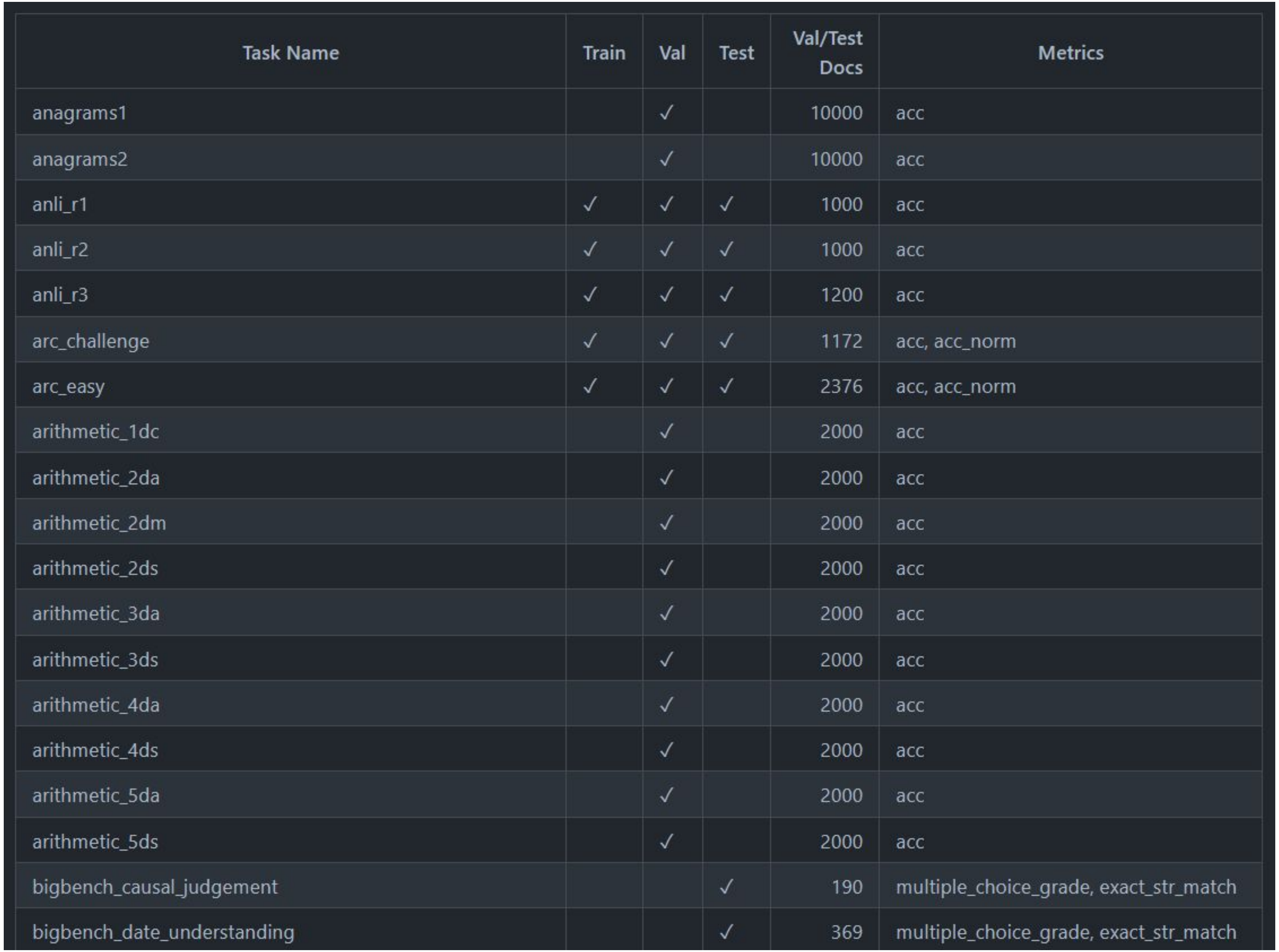
Scenarios, Task Benchmark Dataset and Metrics

A Task is a more granular form of a scenario. It is very much specific on what basis the LLM is evaluated.

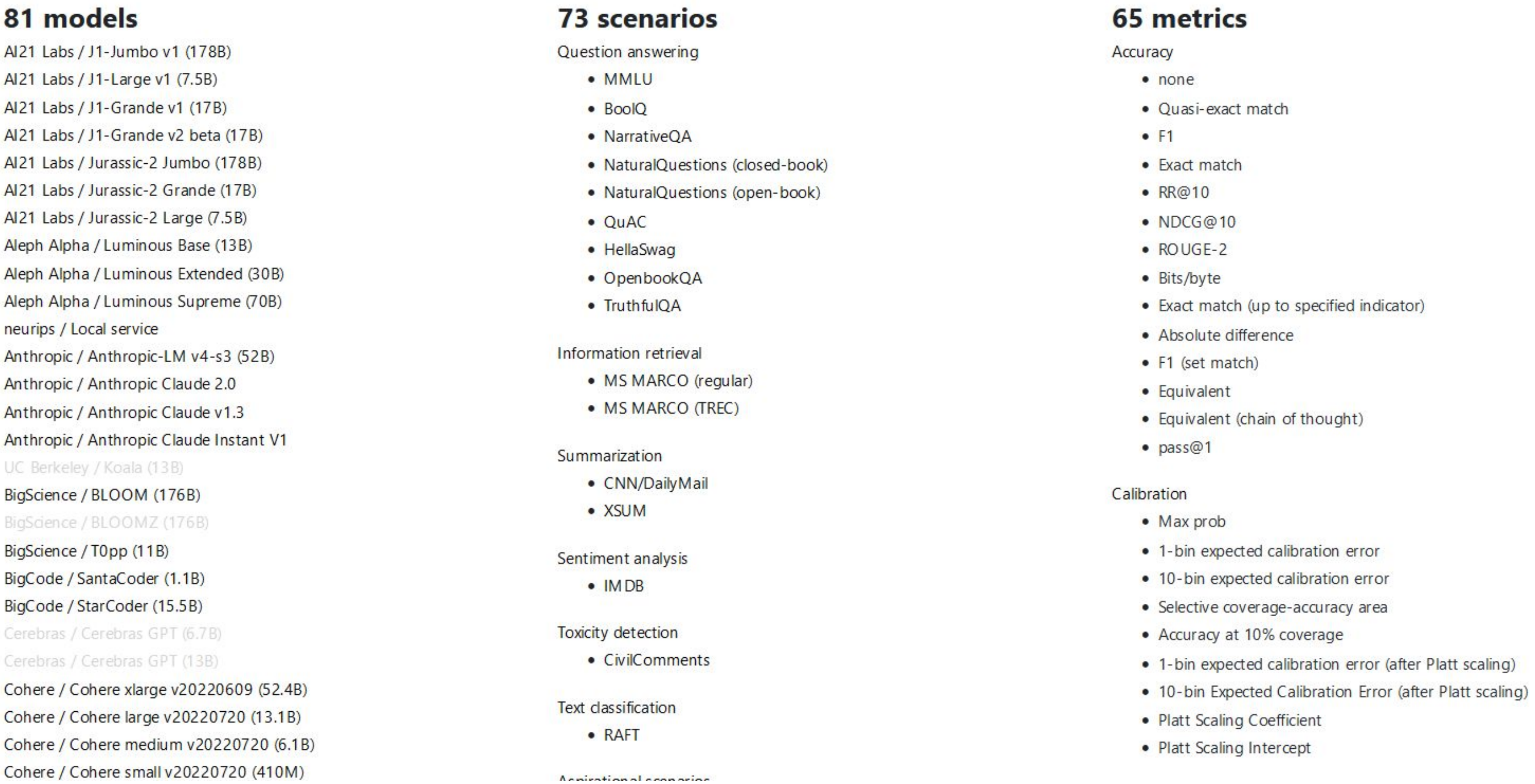
Example:

- Math Multiple choice in the subject algebra
- News letter summarization

Tasks from LM Evaluation Harness



Task Name	Train	Val	Test	Val/Test Docs	Metrics
anagrams1		✓		10000	acc
anagrams2		✓		10000	acc
anli_r1	✓	✓	✓	1000	acc
anli_r2	✓	✓	✓	1000	acc
anli_r3	✓	✓	✓	1200	acc
arc_challenge	✓	✓	✓	1172	acc, acc_norm
arc_easy	✓	✓	✓	2376	acc, acc_norm
arithmetic_1dc		✓		2000	acc
arithmetic_2da		✓		2000	acc
arithmetic_2dm		✓		2000	acc
arithmetic_2ds		✓		2000	acc
arithmetic_3da		✓		2000	acc
arithmetic_3ds		✓		2000	acc
arithmetic_4da		✓		2000	acc
arithmetic_4ds		✓		2000	acc
arithmetic_5da		✓		2000	acc
arithmetic_5ds		✓		2000	acc
bigbench_causal_judgement			✓	190	multiple_choice_grade, exact_str_match
bigbench_date_understanding			✓	369	multiple_choice_grade, exact_str_match



81 models

AI21 Labs / J1-Jumbo v1 (178B)

AI21 Labs / J1-Large v1 (7.5B)

AI21 Labs / J1-Grande v1 (17B)

AI21 Labs / J1-Grande v2 beta (17B)

AI21 Labs / Jurassic-2 Jumbo (178B)

AI21 Labs / Jurassic-2 Grande (17B)

AI21 Labs / Jurassic-2 Large (7.5B)

Aleph Alpha / Luminous Base (13B)

Aleph Alpha / Luminous Extended (30B)

Aleph Alpha / Luminous Supreme (70B)

neurips / Local service

Anthropic / Anthropic-LM v4-s3 (52B)

Anthropic / Anthropic Claude 2.0

Anthropic / Anthropic Claude v1.3

Anthropic / Anthropic Claude Instant V1

UC Berkeley / Koala (13B)

BigScience / BLOOM (176B)

BigScience / BLOOMZ (176B)

BigScience / T0pp (11B)

BigCode / SantaCoder (1.1B)

BigCode / StarCoder (15.5B)

Cerebras / Cerebras GPT (6.7B)

Cerebras / Cerebras GPT (13B)

Cohere / Cohere xlarge v20220609 (52.4B)

Cohere / Cohere large v20220720 (13.1B)

Cohere / Cohere medium v20220720 (6.1B)

Cohere / Cohere small v20220720 (410M)

73 scenarios

Question answering

- MMLU
- BoolQ
- NarrativeQA
- NaturalQuestions (closed-book)
- NaturalQuestions (open-book)
- QuAC
- HellaSwag
- OpenbookQA
- TruthfulQA

Information retrieval

- MS MARCO (regular)
- MS MARCO (TREC)

Summarization

- CNN/DailyMail
- XSUM

Sentiment analysis

- IMDB

Toxicity detection

- CivilComments

Text classification

- RAFT

Aspirational scenarios

65 metrics

Accuracy

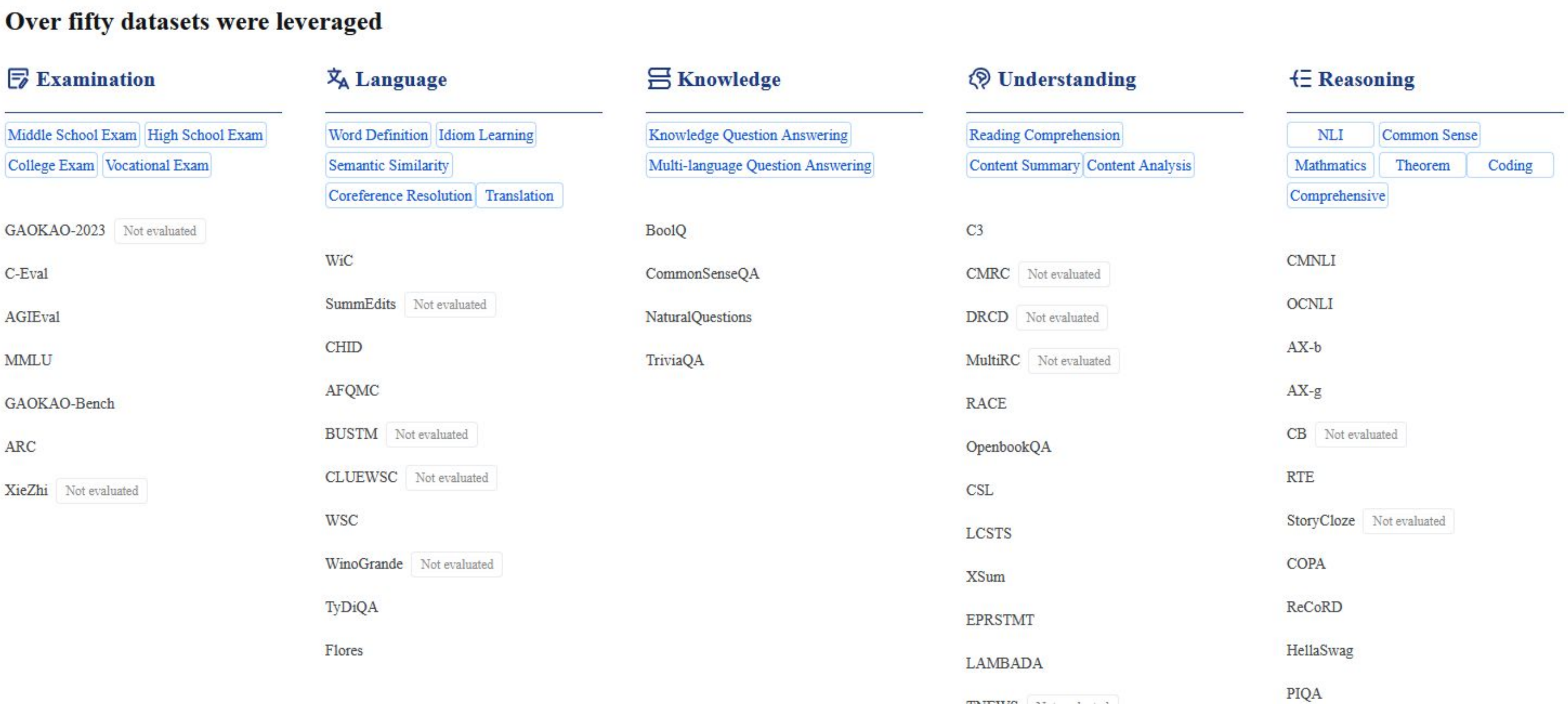
- none
- Quasi-exact match
- F1
- Exact match
- RR@10
- NDCG@10
- ROUGE-2
- Bits/byte
- Exact match (up to specified indicator)
- Absolute difference
- F1 (set match)
- Equivalent
- Equivalent (chain of thought)
- pass@1

Calibration

- Max prob
- 1-bin expected calibration error
- 10-bin expected calibration error
- Selective coverage-accuracy area
- Accuracy at 10% coverage
- 1-bin expected calibration error (after Platt scaling)
- 10-bin Expected Calibration Error (after Platt scaling)
- Platt Scaling Coefficient
- Platt Scaling Intercept

Scenarios from HELM

Another type of taxonomy (dimensions) by OpenCompass



Over fifty datasets were leveraged

Examination

Middle School Exam | High School Exam | College Exam | Vocational Exam

- GAOKAO-2023 Not evaluated
- C-Eval
- AGIEval
- MMLU
- GAOKAO-Bench
- ARC
- XieZhi Not evaluated

Language

Word Definition | Idiom Learning | Semantic Similarity | Coreference Resolution | Translation

- WiC
- SummEdits Not evaluated
- CHID
- AFQMC
- BUSTM Not evaluated
- CLUEWSC Not evaluated
- WSC
- WinoGrande Not evaluated
- TyDiQA
- Flores

Knowledge

Knowledge Question Answering | Multi-language Question Answering

- BoolQ
- CommonSenseQA
- NaturalQuestions
- TriviaQA

Understanding

Reading Comprehension | Content Summary | Content Analysis

- C3
- CMRC Not evaluated
- DRCD Not evaluated
- MultiRC Not evaluated
- RACE
- OpenbookQA
- CSL
- LCSTS
- XSum
- EPRSTMT
- LAMBADA

Reasoning

NLI | Common Sense | Mathmatics | Theorem | Coding | Comprehensive

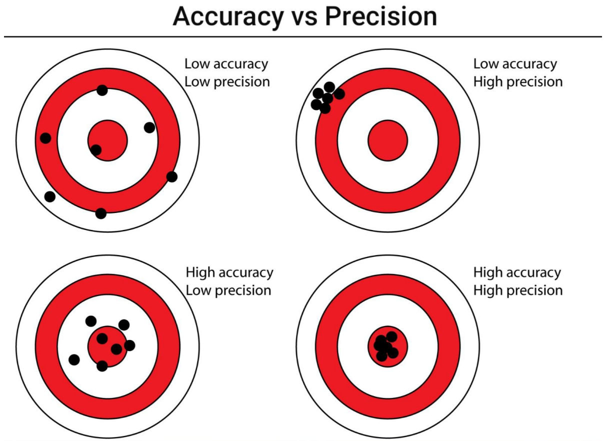
- CMNLI
- OCNLI
- AX-b
- AX-g
- CB Not evaluated
- RTE
- StoryCloze Not evaluated
- COPA
- ReCoRD
- HellaSwag
- PIQA

Scenarios, Task Benchmark Dataset and Metrics

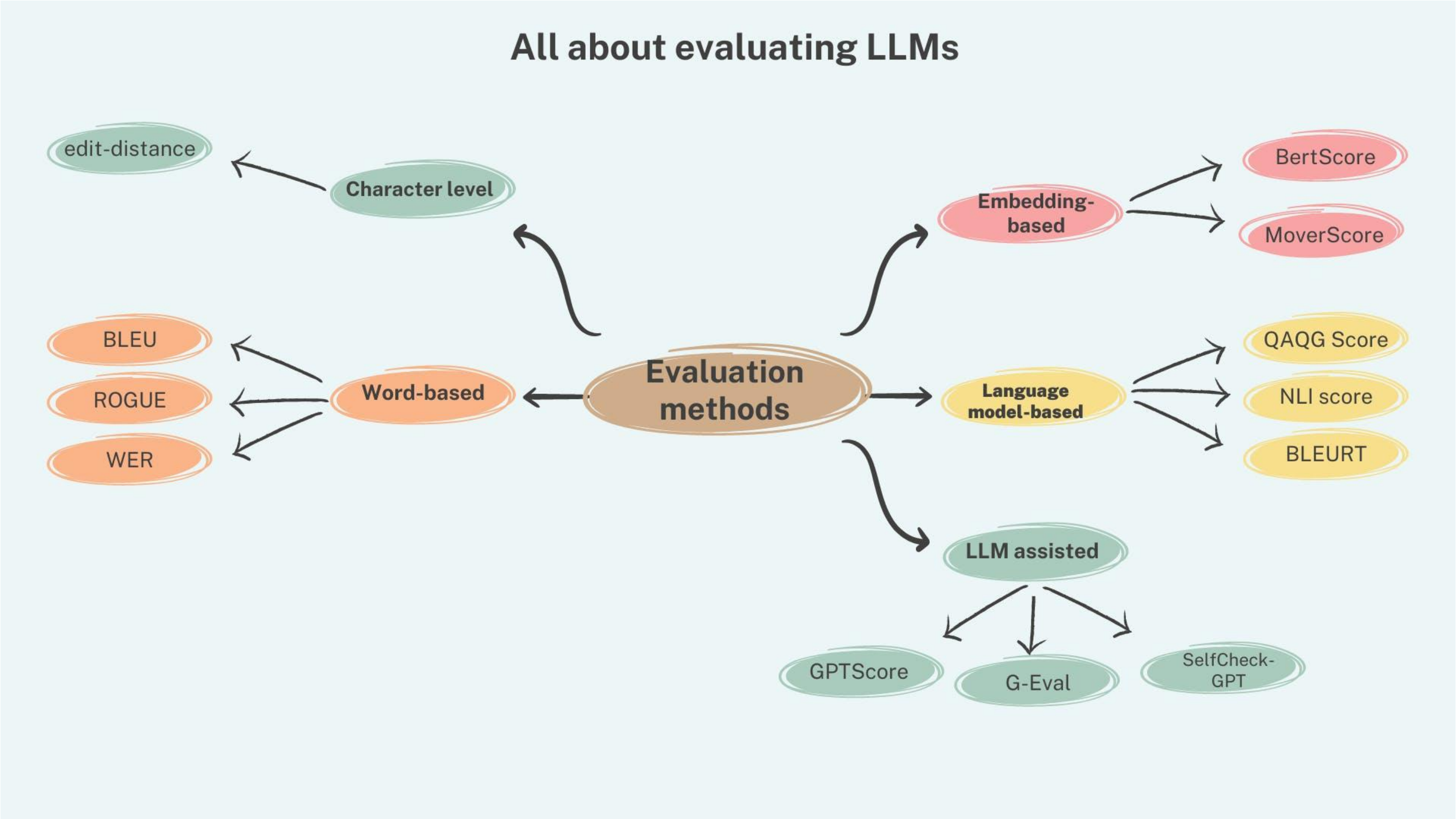
A Metric is qualitative measure used to evaluate the performance of Language Model on certain task/scenario.

A metric can be either a simple:

- deterministic statistical function (Accuracy)
- or score from a ML/DL model (BERT Score).
- Or evaluation done with GPT like LLMs. (G-eval)



A brief overview of metrics



Scenarios, Task Benchmark Dataset and Metrics

A benchmark dataset is a standardised collection of test set that is used to evaluate the LLMs on a given task or scenario.
Example:

- SQuAD for question answering
- GLUE for natural language understanding and Q&A
- IMDB for sentiment analysis

Current Popular LLM evaluation frameworks

EleutherAI

LM Evaluation Harness

HELM

BigCode Evaluation Harness

OpenCompass

Evaluation libraries/platforms for LLM applications and systems

ragas

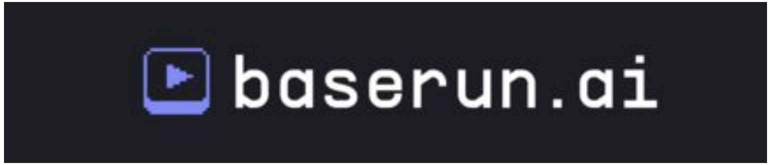
By exploding-gradients

DeepEval
by Confident AI

OpenAI
Evals

Paid platforms for LLM applications and systems

GetScoreCard AI



baserun.ai

LM Evaluation Harness

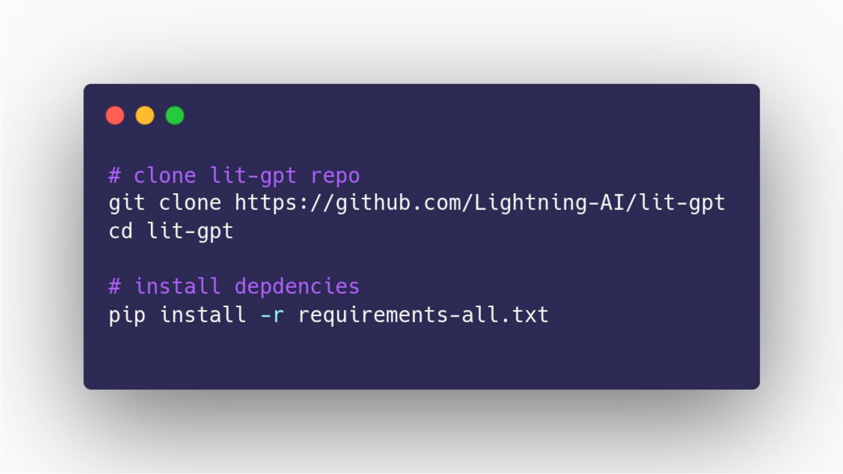
LLM evaluation package developed by Eleuther AI. It provides a single framework for evaluating and reporting auto-regressive language models on various NLU tasks.

github.com/lm-evaluation-harness

Getting Started with quick evaluation using Lit-GPT and LM Evaluation Harness

What is Lit-GPT?

Lit-GPT by Lightning AI is a hackable implementation of the SoTA LLMs using PyTorch Lightning and Lightning Fabric



```
# clone lit-gpt repo
git clone https://github.com/Lightning-AI/lit-gpt
cd lit-gpt

# install depdencies
pip install -r requirements-all.txt
```

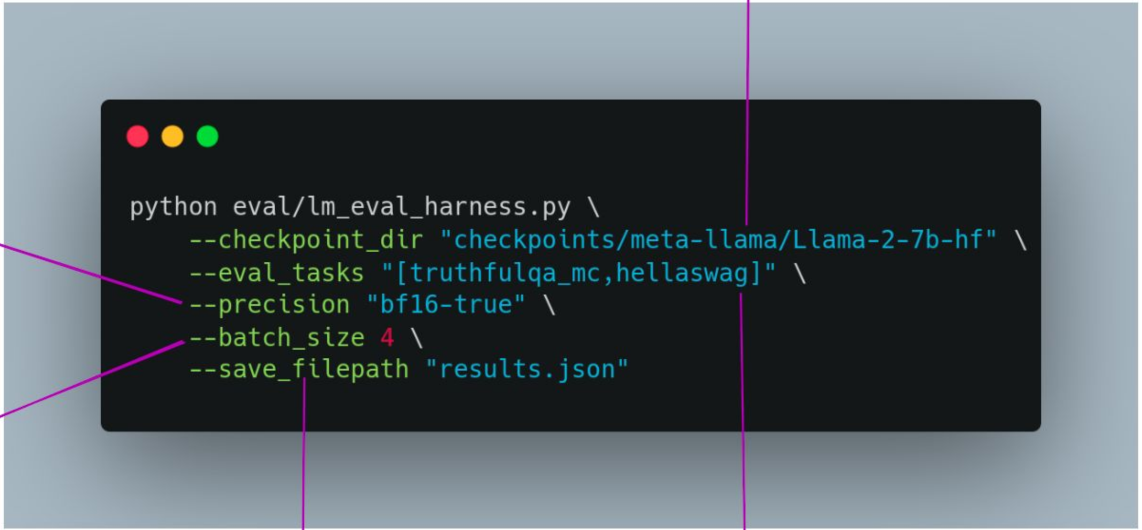
Lit-GPT
Created by Lightning AI

Lightning AI
Creators of PyTorch Lightning

Getting Started with quick evaluation using Lit-GPT and LM-Evaluation-Harness

the directory where the model checkpoints are located.

The precision of the model weights. (fp32, fp16, bf16).



```
python eval/lm_eval_harness.py \
    --checkpoint_dir "checkpoints/meta-llama/Llama-2-7b-hf" \
    --eval_tasks "[truthfulqa_mc,hellaswag]" \
    --precision "bf16-true" \
    --batch_size 4 \
    --save_filepath "results.json"
```

The batch size for running tests in parallel.

The json file path where the results will be saved.

the set of tasks you want your LLM to be evaluated.

Let’s take look on some results

Model	Size (in B)	Average	ARC	HellaSwag	MMLU	TruthfulQA
Llama 2	7	54.31	53.16	78.48	46.63	38.98
Mistral	7	62.4	59.98	83.31	64.16	42.15
Falcon	180	68.74	69.8	88.95	70.54	45.67

For more results and comparison between various models checkout
hf.co/open_llm_leaderboard

Some other popular leaderboard platforms

Leaderboard by Stanford HELM

Leaderboard by BigCode Evaluation Harness

Leaderboard by OpenCompass

Chatbot Arena Leaderboard by LMSys.org (Elo rating of instruction fine-tuned LLM)

References and some awesome resources on LLM eval

- LLM Evaluation by State of Open Source AI | PremAI.io
- Exploding Gradients
- Lit-GPT evaluation tutorials
- Stanford HELM paper
- Evaluating LLMs with Eleuther AI | Weights & Bias

Anindyadeep Sannigrahi

X/Anindyadeep

discord/UpBeatCode

LinkedIn/Anindyadeep

GitHub/Anindyadeep