Artificial Intelligence

Homework #6
Reinforcement Learning with Human Feedback

Outline

1. **Update**
2. **Objective**
3. **Project Overview**
4. **unsloth Explain**
5. **Grading**
6. **Coding**
7. **Report**
8. **Submission**
9. **Useful Links**
10. **Reference**

Update

Thanks for pointing out:

1. In ORPO.py, on line 55, **num_epochs should be changed to num_train_epochs** (NTU COOL files have been updated).
2. The uploaded AI2024-hw6 folder should include "**all**" the **.py** and **.sh** files along with the submission folder, and you should NOT include checkpoints, wandb folder, and outputs folder! (**If you upload checkpoints, other unexpected folders, or are in the wrong format, points will be deducted!**)
3. Extra Experiments: We allow comparisons using models like **chatGPT/Claude/Gemini/etc.,** but please specify the models used (e.g., GPT-4/GPT-4o/GPT-3.5/etc.). **If not specified, we'll deduct points!**

Update (Cont.)

4. 5 ≥ num_train_epochs ≥ 1, If training is less than 1 epoch, we will deduct points!
5. You should also upload a README.md within the AI2024-hw6 folder, for more details please refer to here.
6. Additionally, we provide two smaller LLM options for this assignment, tinyllama-bnb-4bit and gemma-2b-bnb-4bit, for those suffering from the long training time. For more details, please refer to here.
7.

Objective

This assignment aims to utilize the open-source unsloth framework from unslothAI to fine-tune large language models (LLMs) using two reinforcement learning with human feedback (RLHF) techniques, as well as parameter-efficient fine-tuning (PEFT) with LoRA.

made with unsloth

Meta

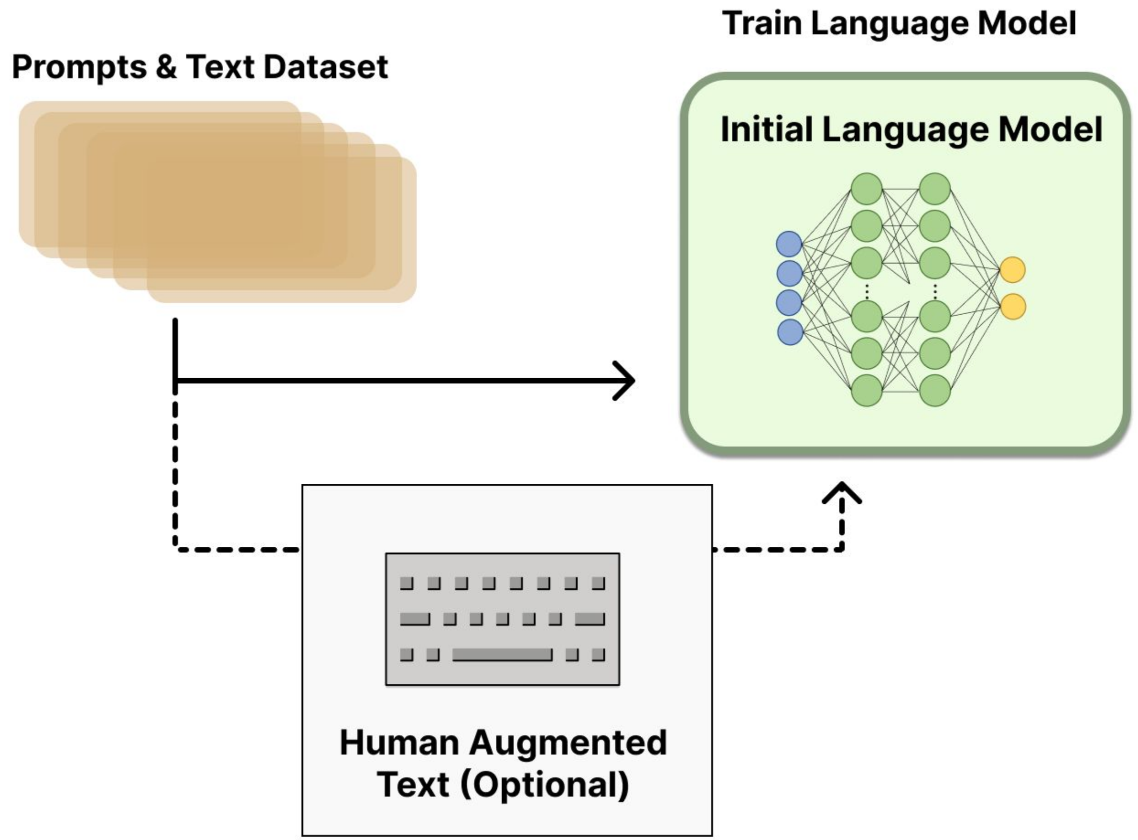
Llama 3

MISTRAL AI_ 7

Objective (Cont.)

Reinforcement Learning with Human Feedback (RLHF)

Step 1:
Supervised Fine-tuning (SFT)

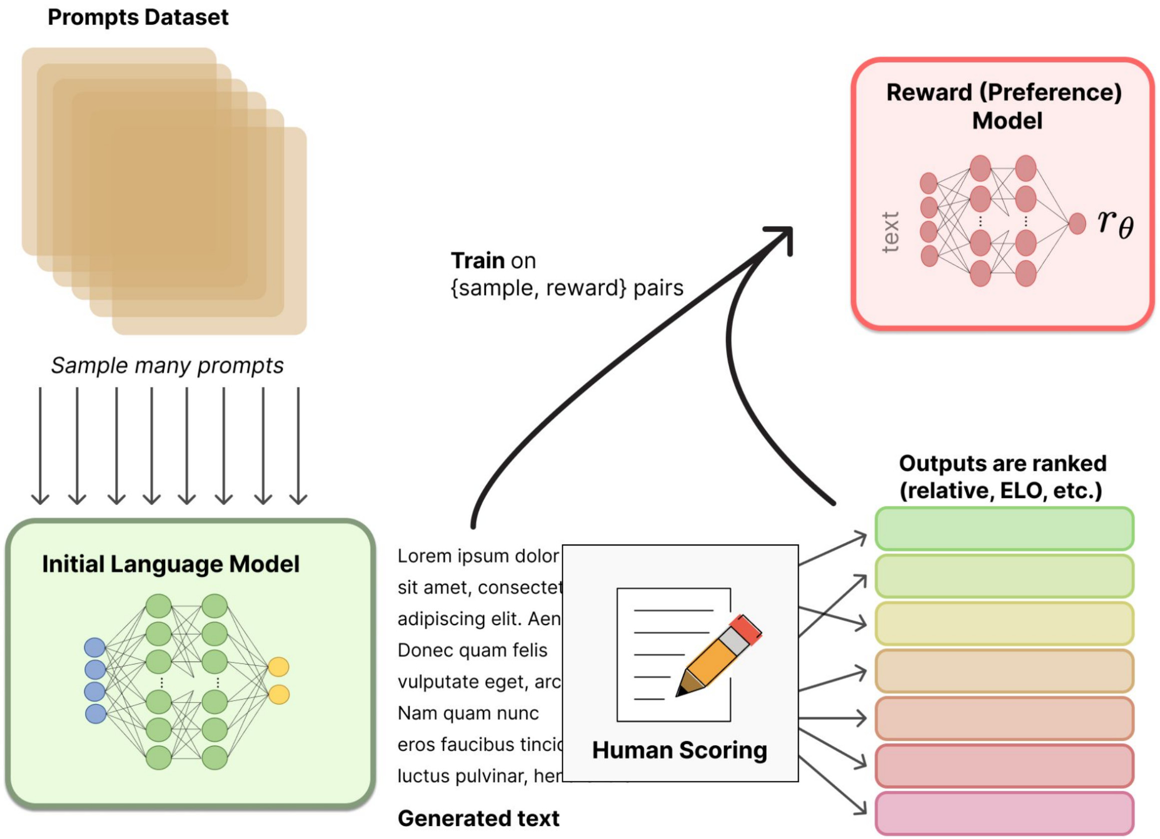


Ref

Objective (Cont.)

Reinforcement Learning with Human Feedback (RLHF)

Step 2: Reward model training

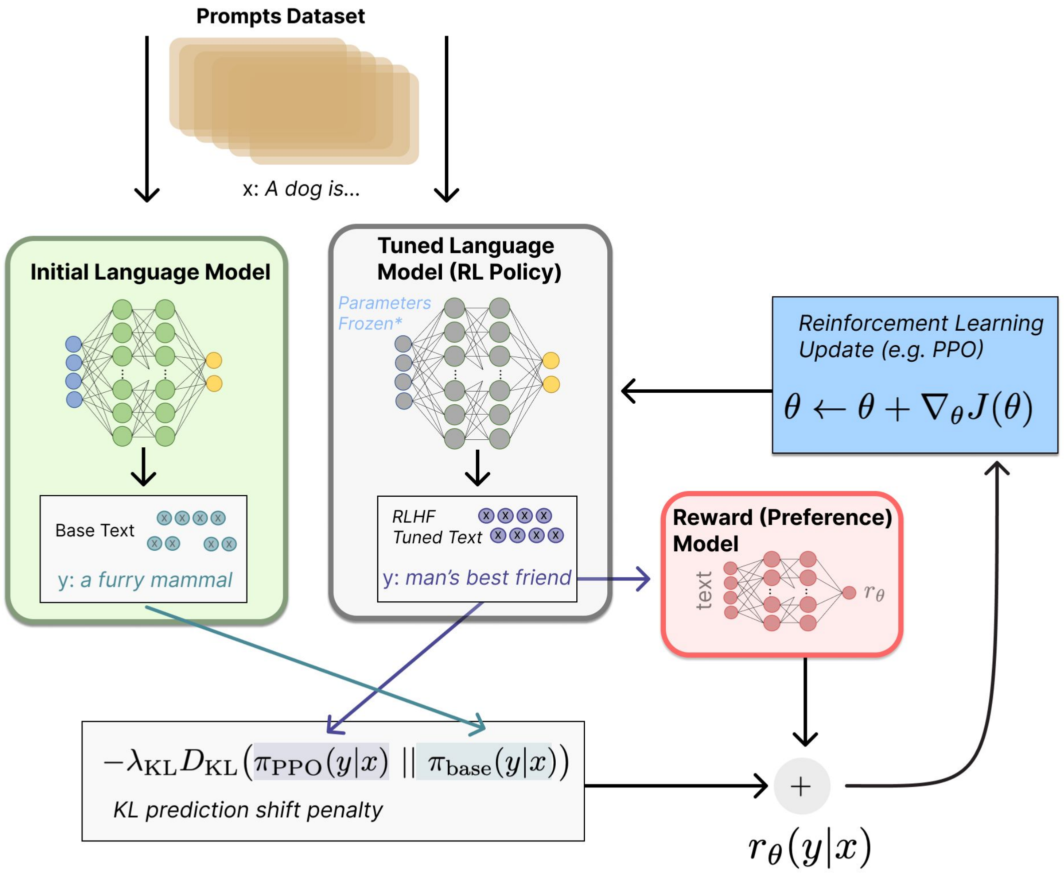


Ref

Objective (Cont.)

Reinforcement Learning with Human Feedback (RLHF)

Step 3: Fine-tuning with RL



Prompts Dataset

x: *A dog is…*

Initial Language Model

Base Text

y: *a furry mammal*

Tuned Language Model (RL Policy)

*Parameters Frozen**

RLHF Tuned Text

y: *man's best friend*

Reinforcement Learning Update (e.g. PPO)

$\theta \leftarrow \theta + \nabla_\theta J(\theta)$

Reward (Preference) Model

text

r_θ

$-\lambda_{\text{KL}} D_{\text{KL}} \left(\pi_{\text{PPO}}(y|x) \,||\, \pi_{\text{base}}(y|x)\right)$

KL prediction shift penalty

$+$

$r_\theta(y|x)$

Ref

Objective (Cont.)

Reinforcement Learning with Human Feedback (RLHF)

Drawbacks:

1. Requires an additional reward model.
2. RL training is unstable and difficult to tune hyperparameters.

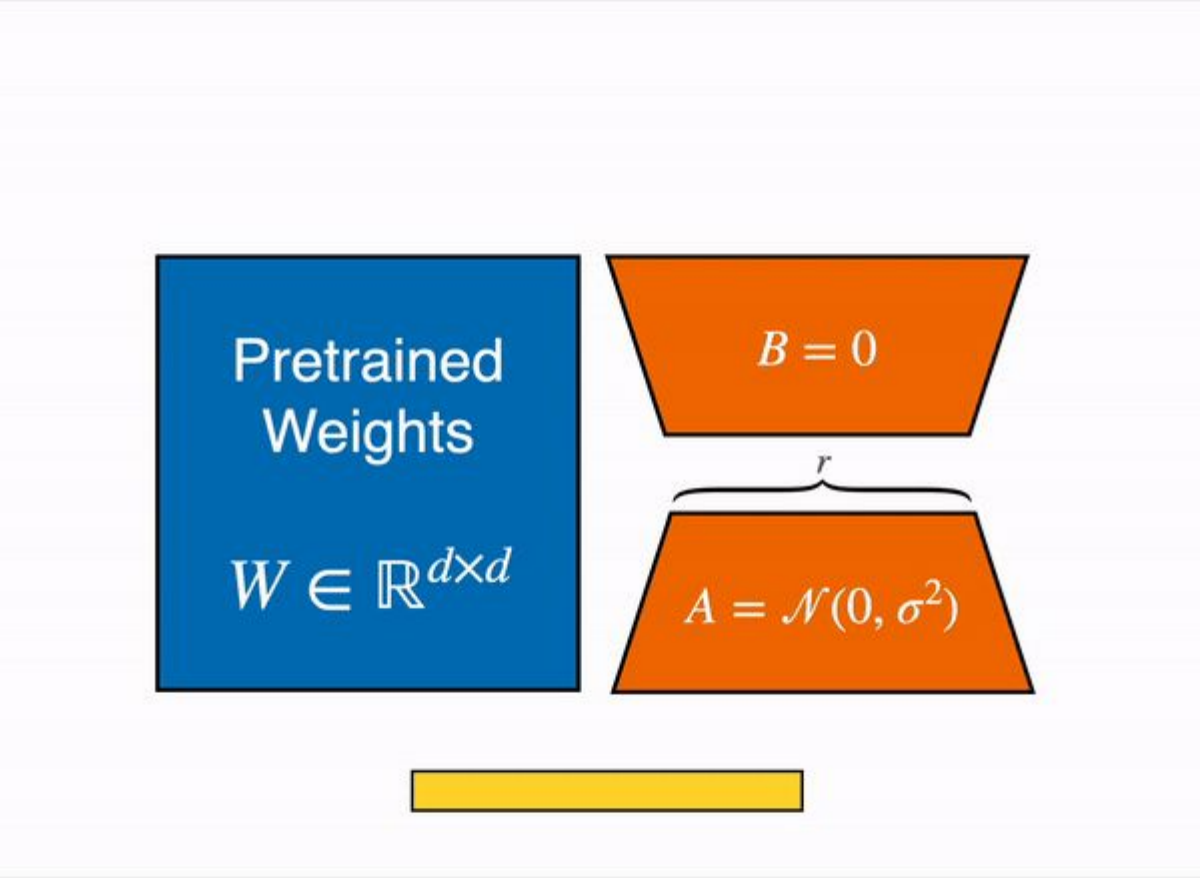
Therefore, we will implement two simplified methods, Direct Preference Optimization (DPO) and Odds Ratio Preference Optimization (ORPO), to fine-tune large language models (LLMs).

Ref

Objective (Cont.)

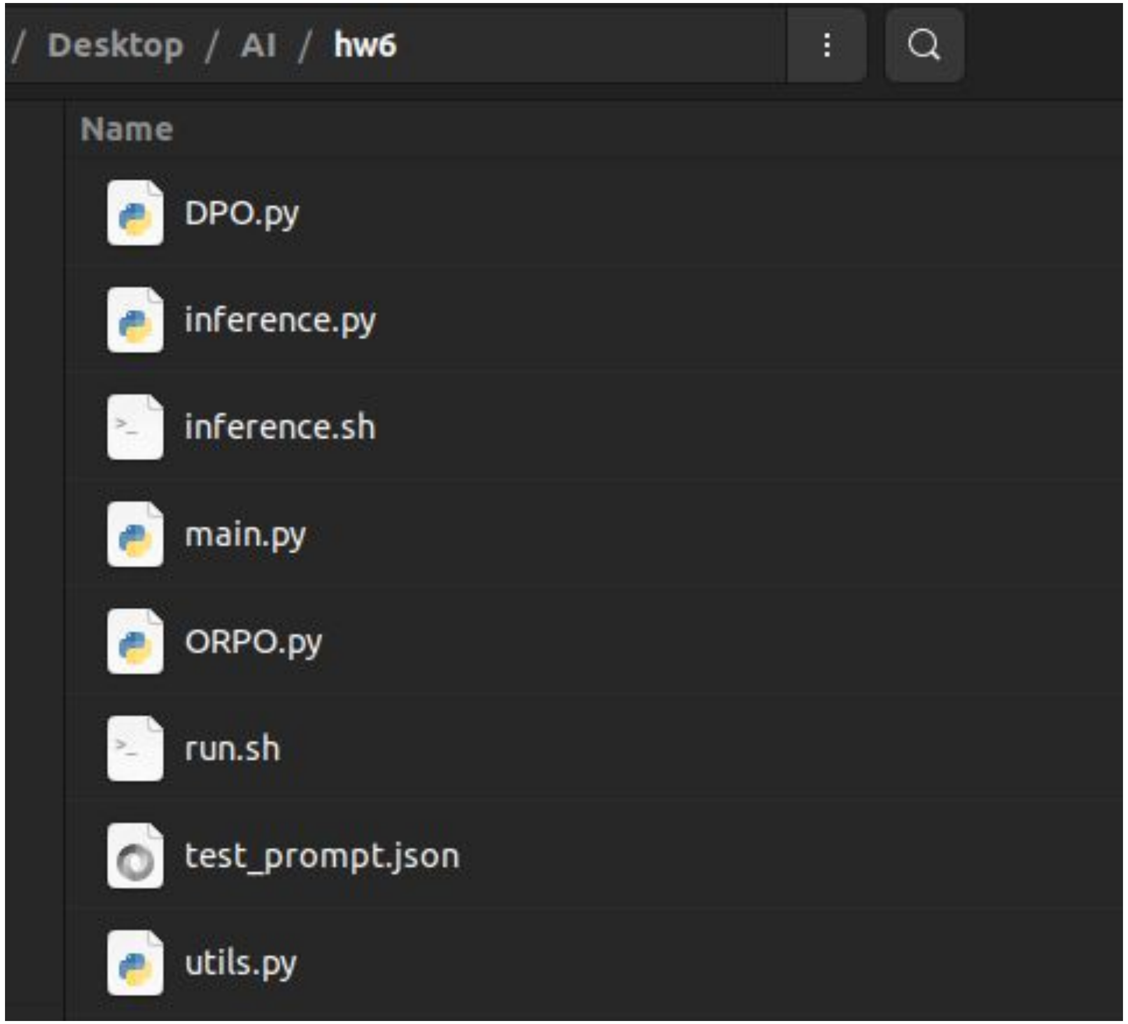
Q: How to fine-tune LLMs on consumer GPU?

A: Low Rank Adaptation (LoRA)


Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$

Ref

Project Overview


/ Desktop / AI / hw6

Name

DPO.py

inference.py

inference.sh

main.py

ORPO.py

run.sh

test_prompt.json

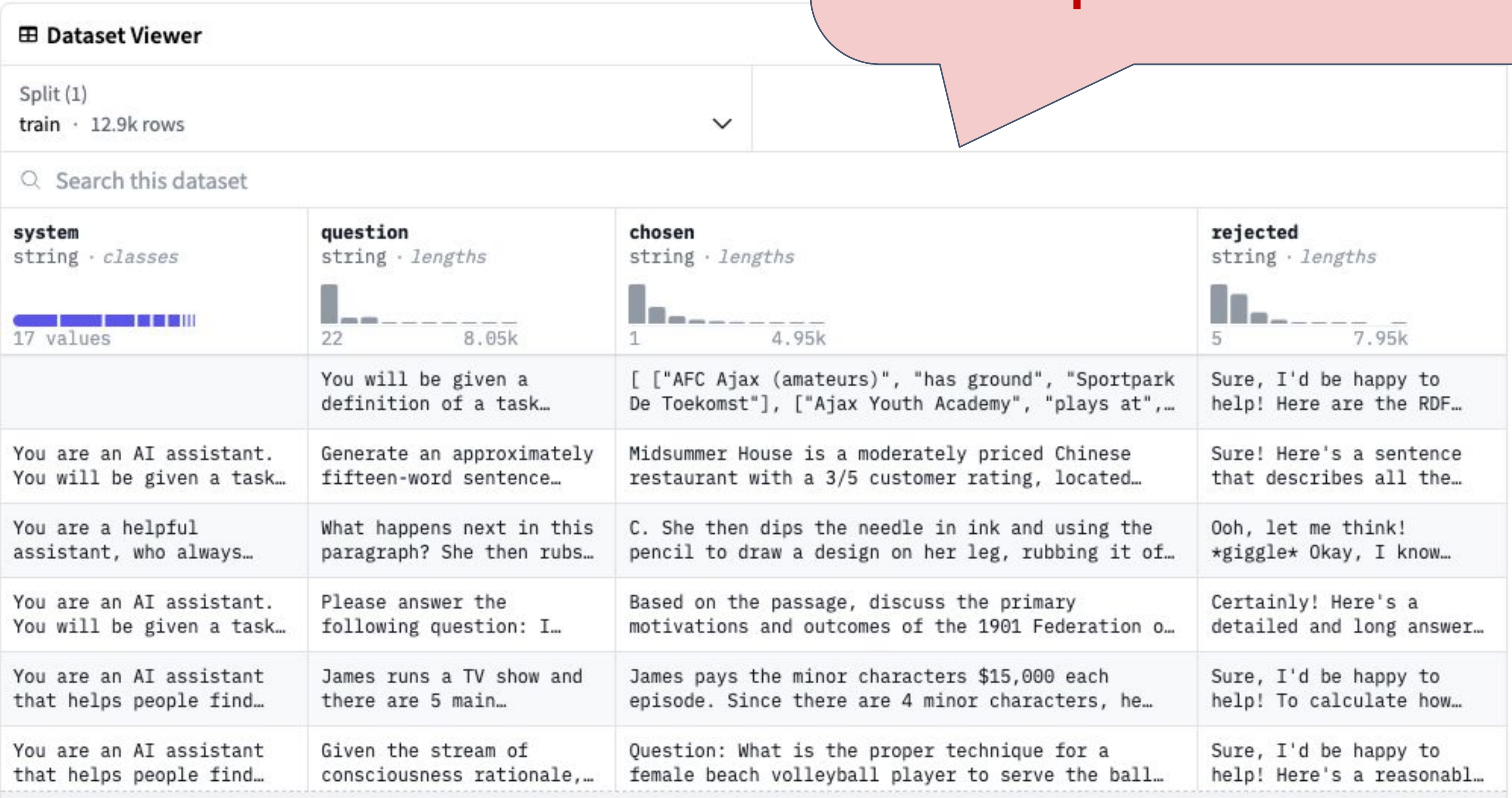
utils.py

- ❖ The code for this project consists of several Python files, **which you will need to read and understand in order to complete the assignment**.
- ❖ You will fill in portions of **DPO.py, and ORPO.py** during the assignment.

Project Overview (Cont.)

Do not change the provided dataset, and using other datasets for data augmentation is not permitted!

Training dataset: Intel/orca_dpo_pairs



Dataset Viewer

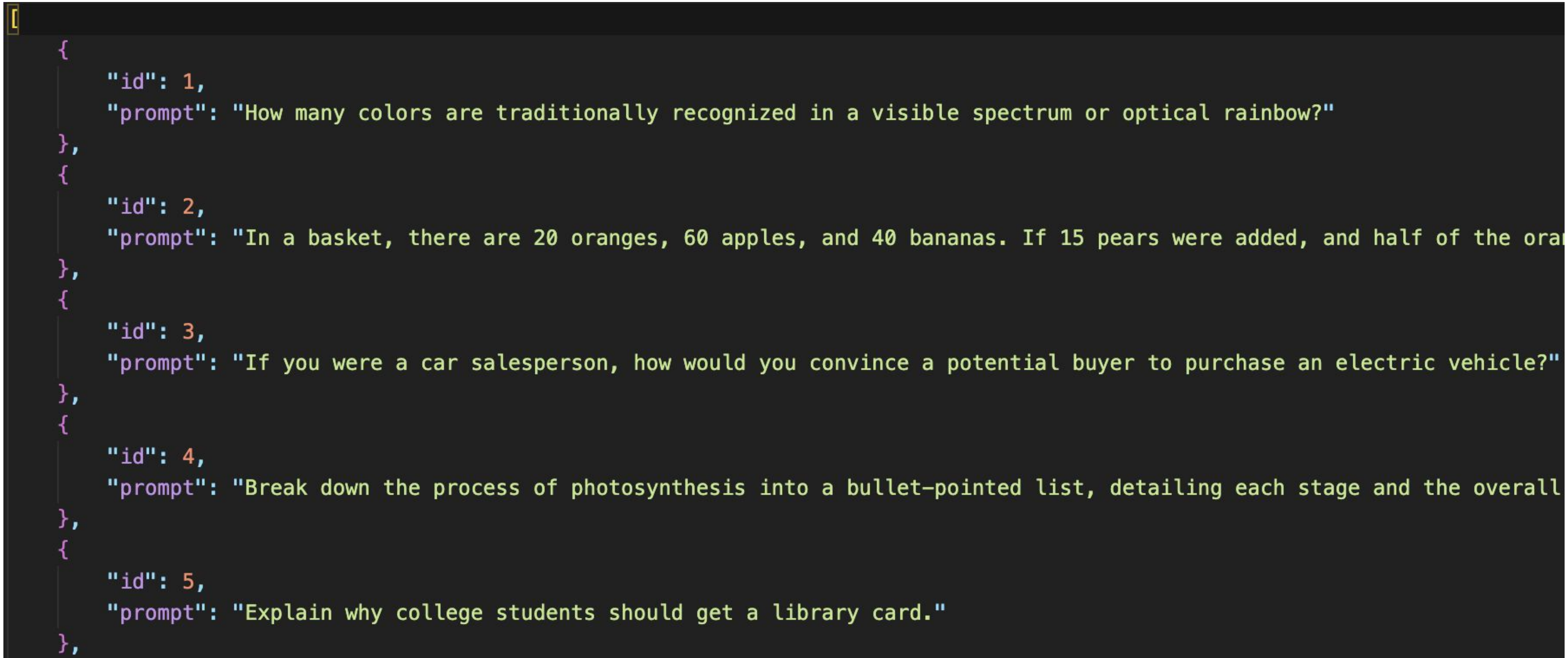
Split (1)
train · 12.9k rows

Search this dataset

system string · *classes* 17 values	question string · *lengths* 22 8.05k	chosen string · *lengths* 1 4.95k	rejected string · *lengths* 5 7.95k
	You will be given a definition of a task…	[["AFC Ajax (amateurs)", "has ground", "Sportpark De Toekomst"], ["Ajax Youth Academy", "plays at",…	Sure, I'd be happy to help! Here are the RDF…
You are an AI assistant. You will be given a task…	Generate an approximately fifteen-word sentence…	Midsummer House is a moderately priced Chinese restaurant with a 3/5 customer rating, located…	Sure! Here's a sentence that describes all the…
You are a helpful assistant, who always…	What happens next in this paragraph? She then rubs…	C. She then dips the needle in ink and using the pencil to draw a design on her leg, rubbing it of…	Ooh, let me think! *giggle* Okay, I know…
You are an AI assistant. You will be given a task…	Please answer the following question: I…	Based on the passage, discuss the primary motivations and outcomes of the 1901 Federation o…	Certainly! Here's a detailed and long answer…
You are an AI assistant that helps people find…	James runs a TV show and there are 5 main…	James pays the minor characters $15,000 each episode. Since there are 4 minor characters, he…	Sure, I'd be happy to help! To calculate how…
You are an AI assistant that helps people find…	Given the stream of consciousness rationale,…	Question: What is the proper technique for a female beach volleyball player to serve the ball…	Sure, I'd be happy to help! Here's a reasonabl…

Project Overview (Cont.)

Testing Dataset: test_prompt.json (released on NTU COOL)



```
[
    {
        "id": 1,
        "prompt": "How many colors are traditionally recognized in a visible spectrum or optical rainbow?"
    },
    {
        "id": 2,
        "prompt": "In a basket, there are 20 oranges, 60 apples, and 40 bananas. If 15 pears were added, and half of the ora
    },
    {
        "id": 3,
        "prompt": "If you were a car salesperson, how would you convince a potential buyer to purchase an electric vehicle?"
    },
    {
        "id": 4,
        "prompt": "Break down the process of photosynthesis into a bullet-pointed list, detailing each stage and the overall
    },
    {
        "id": 5,
        "prompt": "Explain why college students should get a library card."
    },
```

Project Overview (Cont.)

You should choose one of the two required models based on your GPU! However, additional experiments of other models with detailed comparisons are permitted for extra experiments!

LLMs:

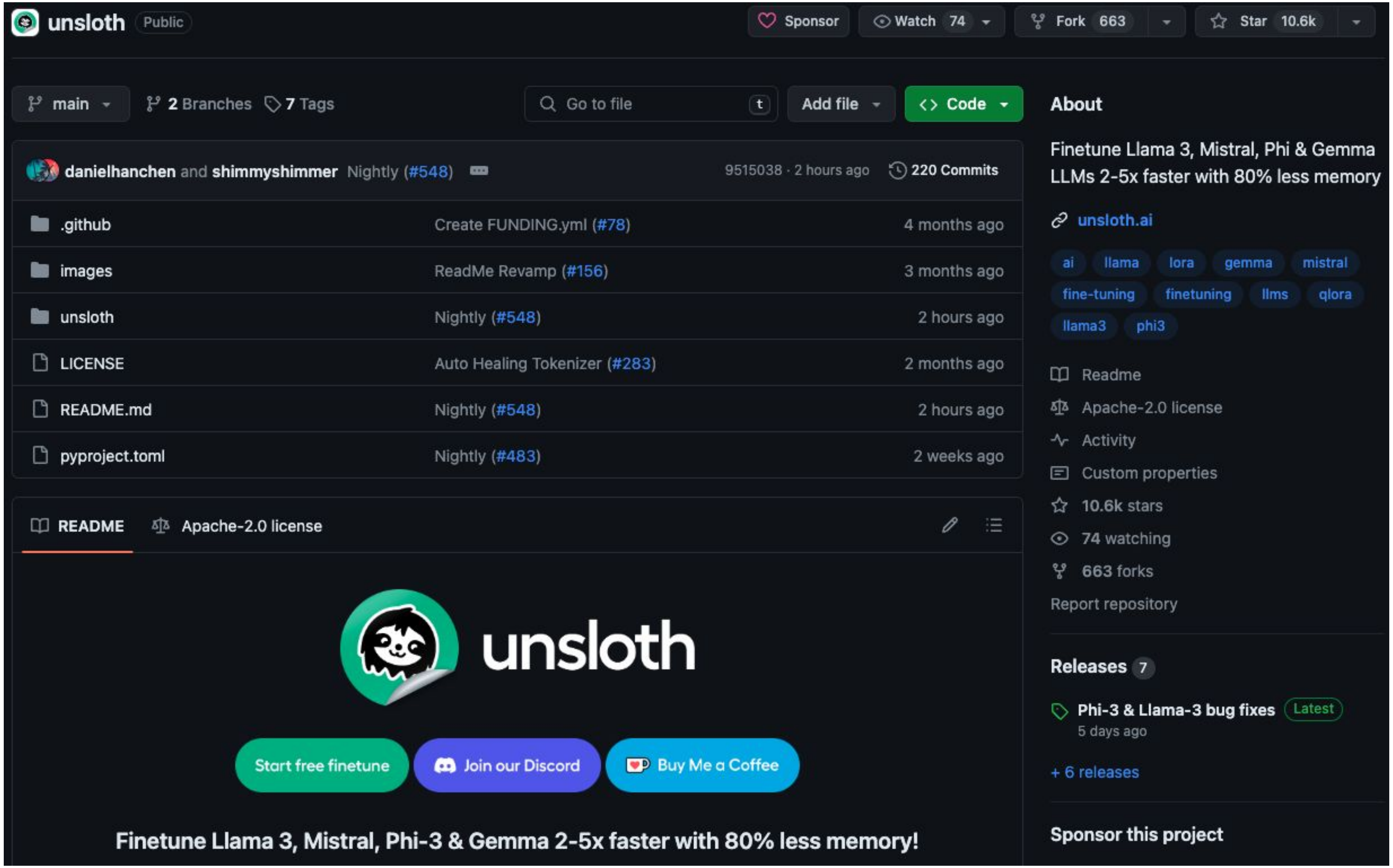
Required (choose one out of four):

1. "unsloth/mistral-7b-v0.3-bnb-4bit"
2. "unsloth/llama-3-8b-bnb-4bit"
3. "unsloth/tinyllama-bnb-4bit"
4. "unsloth/gemma-2b-bnb-4bit"

Additional:

1. "unsloth/mistral-7b-instruct-v0.3-bnb-4bit"
2. "unsloth/llama-3-8b-Instruct-bnb-4bit"
3. …

unsloth Explain



unsloth Public

Sponsor · Watch 74 · Fork 663 · Star 10.6k

main · 2 Branches · 7 Tags · Go to file · Add file · Code

danielhanchen and shimmyshimmer Nightly (#548) · 9515038 · 2 hours ago · 220 Commits

.github	Create FUNDING.yml (#78)	4 months ago
images	ReadMe Revamp (#156)	3 months ago
unsloth	Nightly (#548)	2 hours ago
LICENSE	Auto Healing Tokenizer (#283)	2 months ago
README.md	Nightly (#548)	2 hours ago
pyproject.toml	Nightly (#483)	2 weeks ago

README · Apache-2.0 license

unsloth

Start free finetune · Join our Discord · Buy Me a Coffee

Finetune Llama 3, Mistral, Phi-3 & Gemma 2-5x faster with 80% less memory!

About

Finetune Llama 3, Mistral, Phi & Gemma LLMs 2-5x faster with 80% less memory

unsloth.ai

ai · llama · lora · gemma · mistral · fine-tuning · finetuning · llms · qlora · llama3 · phi3

Readme

Apache-2.0 license

Activity

Custom properties

10.6k stars

74 watching

663 forks

Report repository

Releases 7

Phi-3 & Llama-3 bug fixes Latest
5 days ago

+ 6 releases

Sponsor this project

unsloth Explain (Cont.)

If you are unable to successfully create the virtual environment, it is recommended that you use Colab instead!

P.S. TAs will not assist with troubleshooting environment-related issues.

Installation Instructions:

1. **conda create -y -n ai_hw6 python=3.10**
2. **conda activate ai_hw6**
3. **install pytorch based on your cuda version**
4. **pip install --no-deps trl peft accelerate bitsandbytes**
5. **pip install tqdm packaging wandb**
6. **based on cuda version install the correct version of unsloth (detailed information is here)**
7. **Verify if the installation was successful.:**
   a. **nvcc**
   b. **python -m xformers.info**
   c. **python -m bitsandbytes**

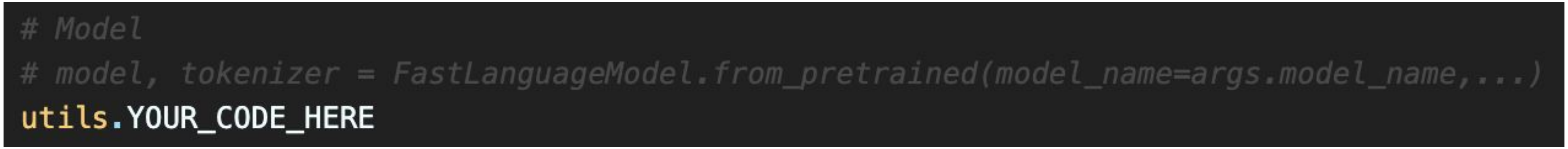
Grading

- **Coding (60%)**
  - DPO.py (20%)
  - ORPO.py (20%)
  - Submission folder (10%)
  - README.md (10%)
- **Report (40%)**
  - Must be submitted in PDF format only.
  - The score for each problem is detailed in here.

Coding (60%)

DPO & ORPO (20%+20%)

- Complete the implementation of DPO and ORPO by filling in the blanks.



```
# Model
# model, tokenizer = FastLanguageModel.from_pretrained(model_name=args.model_name,...)
utils.YOUR_CODE_HERE
```

- Train and inference the fine-tuned models.
- To receive full points, your submitted code should run successfully without any errors.

  (Only one minor modification is allowed for resubmission, but will be subject to a 5% penalty.)
- No partial points will be given.

  Grading: bash run.sh exp_name model_name wandb_token (and other hyperparameters)

Coding (60%) (Cont.)

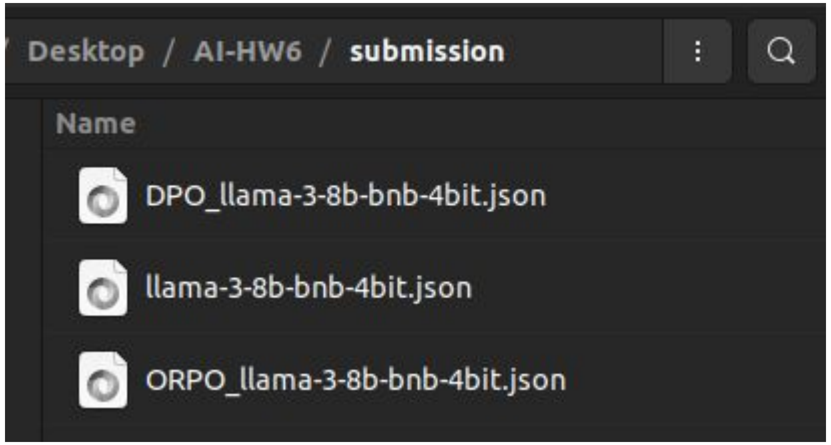
- **Submission folder (10%)**
  - There should be at least **three** JSON files:
    1. the first should contain the generated text from the selected LLM (2%),

       Note: bash inference.sh model_name wandb_token
    2. the second should contain the results after DPO (4%),
    3. and the third should contain the output from ORPO (4%).
  - *Note: submission/*

    -llama-3-8b-bnb-4bit.json

    -DPO_llama-3-8b-bnb-4bit.json

    -ORPO_llama-3-8b-bnb-4bit.json



DPO_llama-3-8b-bnb-4bit.json

llama-3-8b-bnb-4bit.json

ORPO_llama-3-8b-bnb-4bit.json

(If you have done extra experiments, you must also include those files from different hyperparameters or models in the submission directory!)

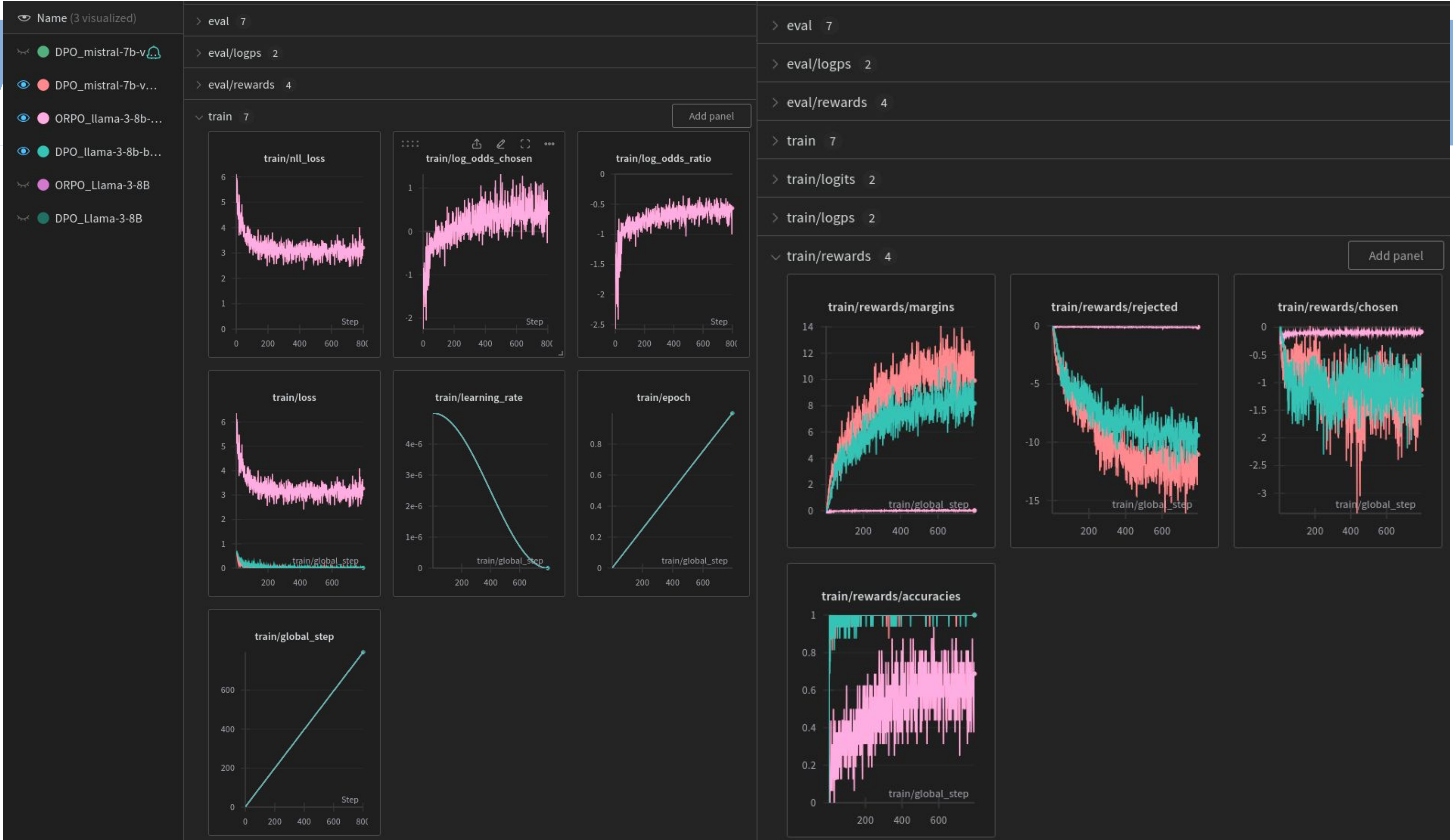
Coding (60%) (Cont.)

- **README.md (10%)**
  - In the README.md, you should provide following information:
    1. commands for creating the virtual env, thus **requirements.txt is required if specified in the README.md!**
    2. how to train your model, including "**all**" the hyperparameters you used should be specified. (feel free to modify the run.sh)

(**We'll follow the steps provided in your README.md for reproducibility. If we can not successfully run your code, we will then deduct points! Thus, make sure your README.md is detailed enough before submission!**)

Report (40%)

- **Report (25%)**
  - Provide a brief description and comparison of DPO and ORPO. (5%+5%)
  - Briefly describe LoRA. (5%)
  - Plot your training curve by W&B, including both loss and rewards. (5%)
  - Comparison and analysis of results (before & after DPO & after ORPO) (5%)


Name (3 visualized)
DPO_mistral-7b-v
DPO_mistral-7b-v...
ORPO_llama-3-8b-...
DPO_llama-3-8b-b...
ORPO_Llama-3-8B
DPO_Llama-3-8B
eval 7
eval/logps 2
eval/rewards 4
train 7
Add panel
train/nll_loss
0
1
2
3
4
5
6
Step
0
200
400
600
800
train/log_odds_chosen
1
0
-1
-2
Step
0
200
400
600
800
train/log_odds_ratio
0
-0.5
-1
-1.5
-2
-2.5
Step
0
200
400
600
800
train/loss
0
1
2
3
4
5
6
train/global_step
200
400
600
train/learning_rate
0
1e-6
2e-6
3e-6
4e-6
train/global_step
200
400
600
train/epoch
0
0.2
0.4
0.6
0.8
train/global_step
200
400
600
train/global_step
0
200
400
600
Step
0
200
400
600
800
eval 7
eval/logps 2
eval/rewards 4
train 7
train/logits 2
train/logps 2
train/rewards 4
Add panel
train/rewards/margins
0
2
4
6
8
10
12
14
train/global_step
200
400
600
train/rewards/rejected
0
-5
-10
-15
train/global_step
200
400
600
train/rewards/chosen
0
-0.5
-1
-1.5
-2
-2.5
-3
train/global_step
200
400
600
train/rewards/accuracies
0
0.2
0.4
0.6
0.8
1
train/global_step
200
400
600

Report (40%) (Cont.)

- **Extra Experiments (15%) – <u>We will assign grades based on the richness of the experiments, with different score regarding to varying levels of detail and complexity.</u>**
  - Comparison and analysis of various hyperparameters (e.g., num_epochs, beta, etc.)
  - Comparison and analysis of various models, which GPT-4, GPT-4o, GPT-3.5, and etc. are allowed, for the variants of the required models please refer to here.

  (If you have done extra experiments, you **must** also include those files from different hyperparameters or models in the submission directory!)

Submission

- Deadline: 2024/06/12 (Wed.) 23:59
- Zip all files as **hw6_<student_id>.zip**
- Submit to NTU COOL
- Your submission should include the following files:
  - hw6_<student_id>.pdf
  - AI2024-hw6
- **Do not put report.pdf into AI2024-hw6 folder**
- *Note: hw6_<student_id> is an example format. For instance, if your student ID is r1234567, then the file name should be hw6_r1234567 or hw6_R1234567.*

Late Policy

- The discount for late submission of assignments is as follows:

<24hr	<48hr	<72hr
70%	50%	30%

Useful Links

RLHF:

https://www.youtube.com/watch?v=v12IKvF6Cj8

https://speech.ee.ntu.edu.tw/~hylee/genai/2024-spring-course-data/0412/0412_LLMtraining_part3.pdf

W&B:

https://docs.wandb.ai/quickstart

https://docs.wandb.ai/tutorials

Reference

https://huggingface.co/blog/rlhf

https://huggingface.co/blog/trl-peft

https://hackmd.io/@YungHuiHsu/Sy5Ug7iV6

Any Question

ai.ta.2024.spring@gmail.com