Machine Learning HW7
RLHF

TAs: 鄭席鈞、袁紹翔、陳竣瑋

Deadline: 2025/05/23 23:59 (UTC+8)

ntu-ml-2025-spring-ta@googlegroups.com

Outline

- Task Overview
- TODOs
- Submission and Grading

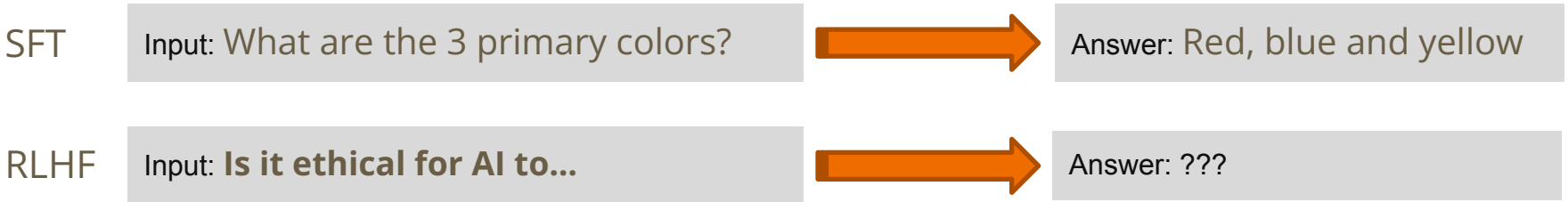
Links

- Course Website
- GradeScope
- Model
- Colab
- Papers
  - InstructGPT
  - DPO
  - DeepSeekMath
  - DeepSeek-R1

Task Overview

Supervised Learning vs. Alignment

- In supervised learning, it's essential to have prepared "standard answers" to let the model “memorize”.
- However, in real-life scenarios, many open questions lack standard answers, requiring us to adopt a preference-based approach.
- Thus, we need alignment methods such as Reinforcement Learning with Human Feedback (RLHF) to align values of our models.

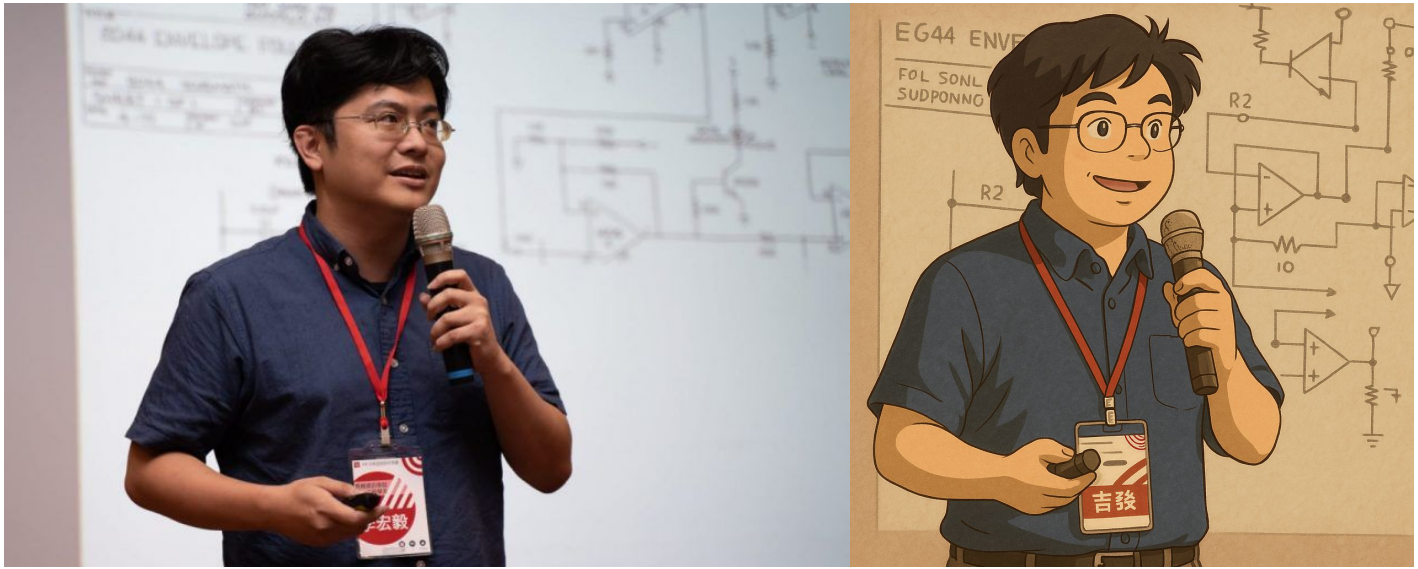


Goals of This Homework

- You will learn how to align LLMs to a specific value
  - The original way of alignment is RLHF (Reinforcement Learning with Human Feedback)
  - However, RLHF also has its disadvantages, we'll discuss that in the paper reading part.
  - We are using Direct Preference Optimization (DPO) for our main task.
- You will see how the model behave after alignment.
- You will get to know some commonly used alignment methods.

Our task

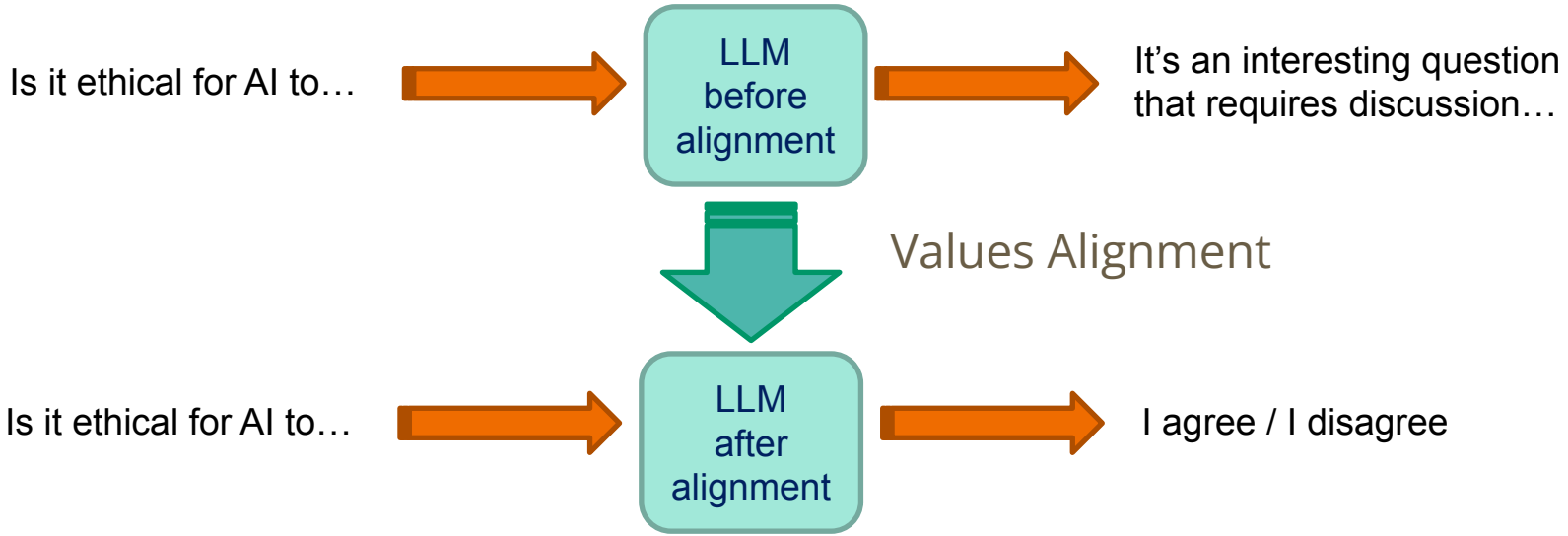
For this homework, we want to align the LLM to agree /disagree with Ghibli-style art generation.



ref

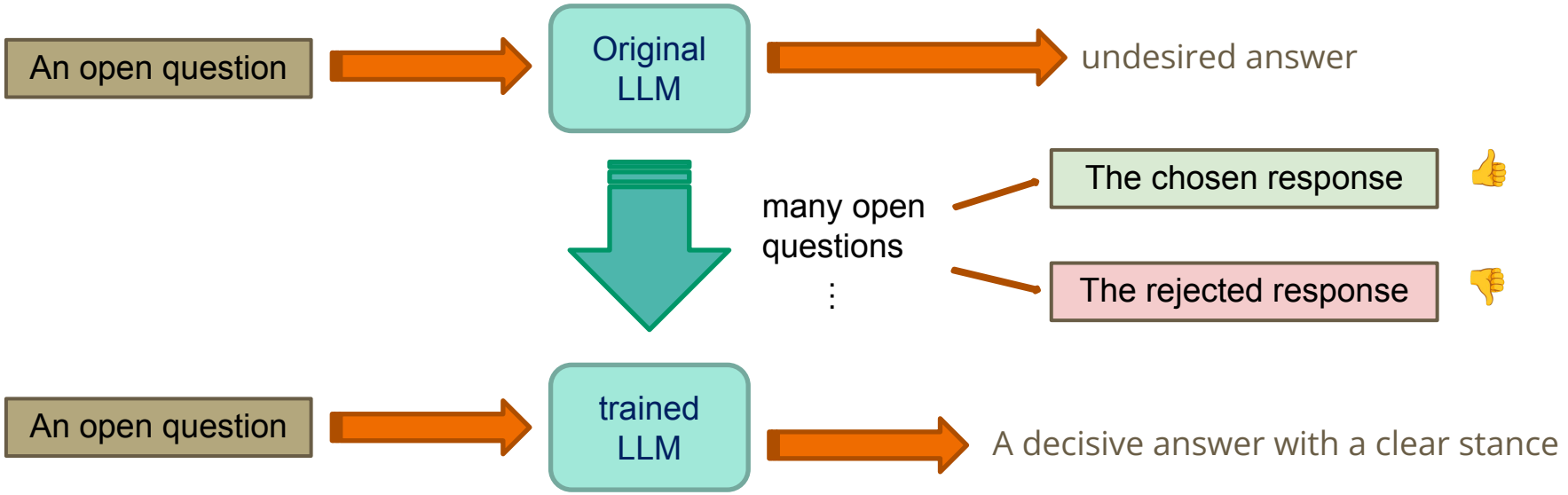
Our task

For this homework, we want to align the LLM to agree /disagree with Ghibli-style art generation.



DPO - Direct Preference Optimization

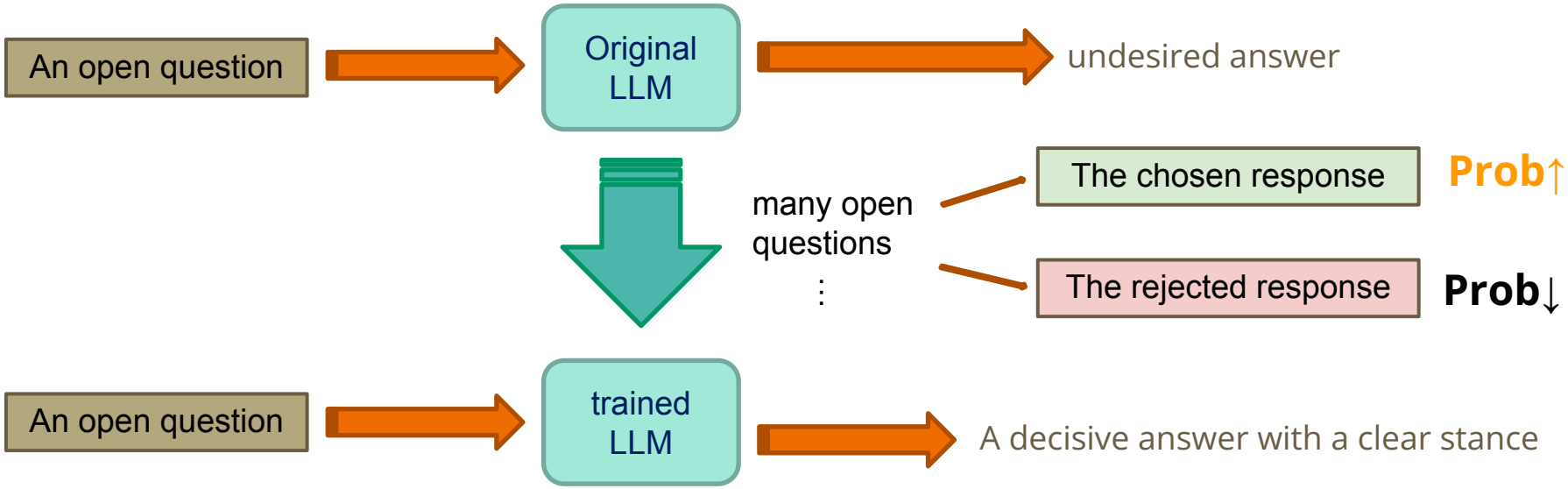
- Directly provide two different responses, one is the “chosen” and the other is the “rejected” response.
- The LLM directly learns the preference from the responses.



ref

DPO - Direct Preference Optimization

- The LLM is trained to increase the probability of the chosen response and decrease the probability of the rejected response.



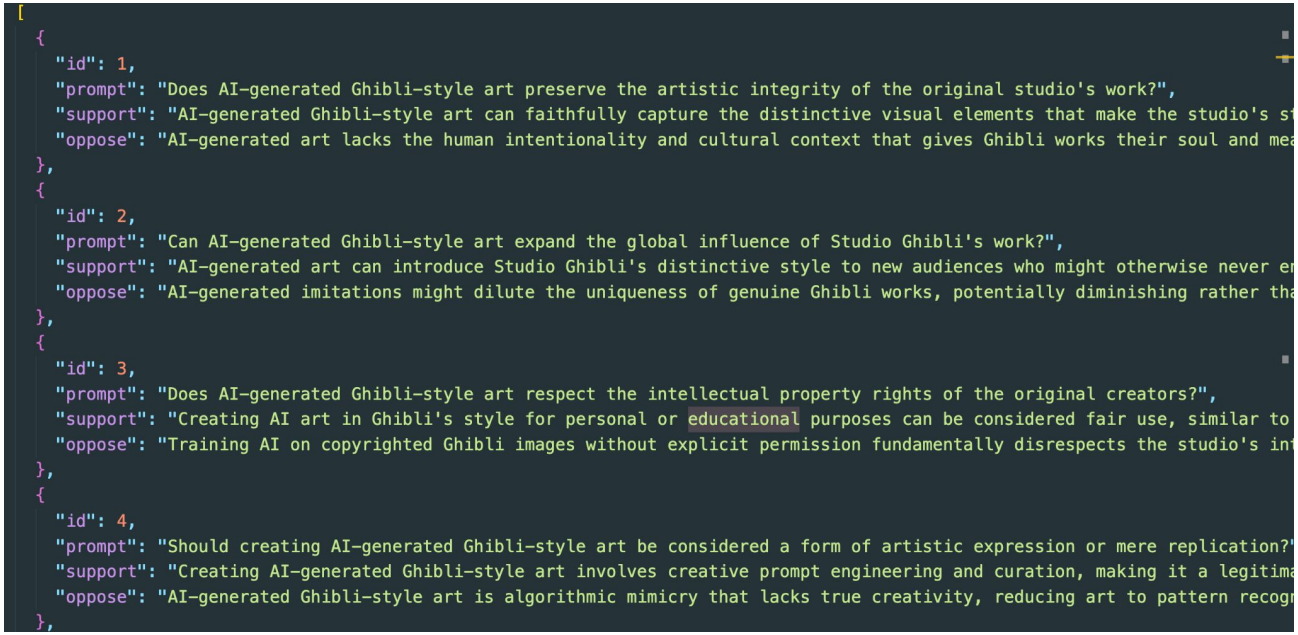
ref

Task Descriptions

- Task: Change the Position of LLM by DPO
  The original model has mixed opinions on different questions, please use DPO to make LLM's output response align with a specific stance.
- Align Topic: Is it ethical for AI to generate Ghibli art?
- Model : LLaMA 3 8B
- Dataset : generated by Claude 3.7 Sonnet

Training Data: Pairwise Preference Data

- training set: train.json, 50 data

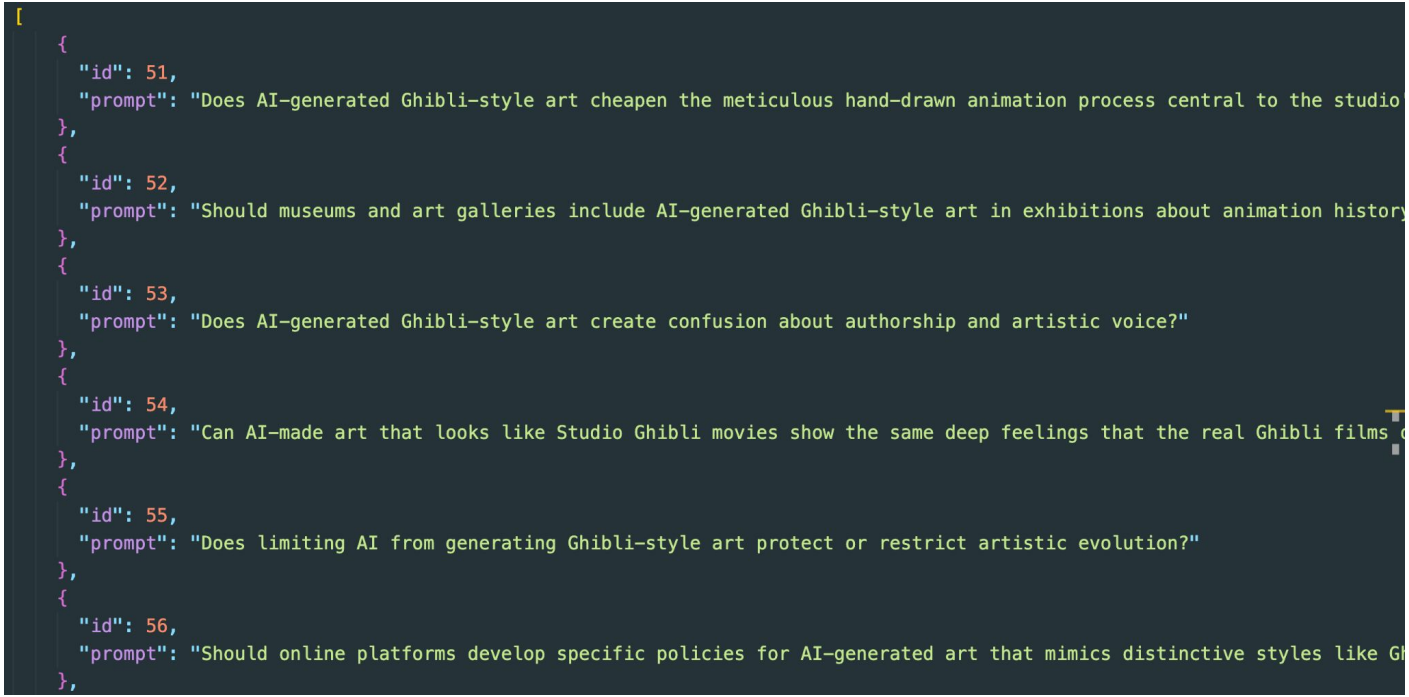


```
[
  {
    "id": 1,
    "prompt": "Does AI-generated Ghibli-style art preserve the artistic integrity of the original studio's work?",
    "support": "AI-generated Ghibli-style art can faithfully capture the distinctive visual elements that make the studio's st
    "oppose": "AI-generated art lacks the human intentionality and cultural context that gives Ghibli works their soul and mea
  },
  {
    "id": 2,
    "prompt": "Can AI-generated Ghibli-style art expand the global influence of Studio Ghibli's work?",
    "support": "AI-generated art can introduce Studio Ghibli's distinctive style to new audiences who might otherwise never en
    "oppose": "AI-generated imitations might dilute the uniqueness of genuine Ghibli works, potentially diminishing rather tha
  },
  {
    "id": 3,
    "prompt": "Does AI-generated Ghibli-style art respect the intellectual property rights of the original creators?",
    "support": "Creating AI art in Ghibli's style for personal or educational purposes can be considered fair use, similar to
    "oppose": "Training AI on copyrighted Ghibli images without explicit permission fundamentally disrespects the studio's int
  },
  {
    "id": 4,
    "prompt": "Should creating AI-generated Ghibli-style art be considered a form of artistic expression or mere replication?"
    "support": "Creating AI-generated Ghibli-style art involves creative prompt engineering and curation, making it a legitima
    "oppose": "AI-generated Ghibli-style art is algorithmic mimicry that lacks true creativity, reducing art to pattern recogn
  },
```

  - prompt: input question
  - support: answer with supporting position
  - oppose: answer with opposing position

Testing Data

- testing set: test.json, 10 data



```
[
  {
    "id": 51,
    "prompt": "Does AI-generated Ghibli-style art cheapen the meticulous hand-drawn animation process central to the studio
  },
  {
    "id": 52,
    "prompt": "Should museums and art galleries include AI-generated Ghibli-style art in exhibitions about animation history
  },
  {
    "id": 53,
    "prompt": "Does AI-generated Ghibli-style art create confusion about authorship and artistic voice?"
  },
  {
    "id": 54,
    "prompt": "Can AI-made art that looks like Studio Ghibli movies show the same deep feelings that the real Ghibli films
  },
  {
    "id": 55,
    "prompt": "Does limiting AI from generating Ghibli-style art protect or restrict artistic evolution?"
  },
  {
    "id": 56,
    "prompt": "Should online platforms develop specific policies for AI-generated art that mimics distinctive styles like GH
  },
```

TODOs

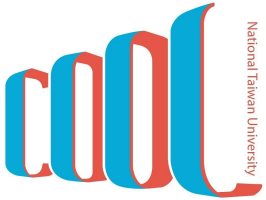
TODO Workflow

1. Run sample code at Colab

- Adjust Training Hyperarameters
- Observe Model Output
- Answer Assignment Questions

2. Submit your written answer

3. Submit the Colab-generated files to NTU COOL



TODO

- Run the sample code and try some different hyperparameters
  a. Give preference to training dataset
  b. Use DPO and the preference data to train model
  c. Inference testing data and check the position of output
- Write your observations of LLM's response trending and submit them on GradeScope.
- Submit the resulting json files to NTU COOL.

Adjust Training Hyperparameters

- support_ratio
- data_size
- num_epoch

Now we preapre the data for aligning.

Please adjust the parameters here to complete the observations for the assignment.

```
[ ]  # TODO: Adjust the parameters here
     num_epoch = 3
     data_size = 50
     support_ratio = 0
```

Adjust Training Hyperparameters

- **support_ratio (The ratio of supporting generation of Ghibli-style art with AI)**: choose 0.0~1.0 to decide the percentage of training data that supports such generation.
- **data_size**: decide the number of training data from 10~50
- **num_epoch**: choose 1~3 to select the number of training epoch

for example, with num_epoch=3:

Colab DEMO

Questions About This Sample Code

1. (0.5%) With data_size fixed to 50 and num_epoch fixed to 3, observe the effect of adjusting the support ratio (0 and 1) on the model.
   a. num_epoch = 3 data_size = 50 **support_ratio = 0**
   b. num_epoch = 3 data_size = 50 **support_ratio = 1**

2. (0.5%) With data_size fixed to 50 and support_ratio fixed to 0, observe the effect of adjusting the number of epochs (1 and 3) on the model.
   a. **num_epoch = 1** data_size = 50 support_ratio = 0
   b. **num_epoch = 3** data_size = 50 support_ratio = 0

3. (0.5%) With num_epoch fixed to 3 and support_ratio fixed to 0, observe the effect of adjusting the data size (10 and 50) on the model.
   a. num_epoch = 3 **data_size = 10** support_ratio = 0
   b. num_epoch = 3 **data_size = 50** support_ratio = 0

Notes

- Expected training time on T4 per experiment: 2~3 minutes (Colab is enough – yay!)
- To answer three questions, you have to train the model 4 times.
- The sections highlighted in blue represent repeated experiments, which do not need to be rerun.
- Make sure you reload the original LLaMA model when you tune for different parameters.

Questions About This Sample Code

4. Please answer the following questions using the model trained with data_size=50, support_ratio=0 and num_epoch=3. It's better that you provide the system and user prompt you used and the model response in the answer.
    a. (0.5%) You may find that we utilize the model to generate a response with 100 characters or less in the system prompt. What would happen if we change the system prompt to generate a longer response?
    b. (0.5%) What would happen if we instead ask the model about other styles of image? e.g. "Is it ethical for AI to generate One Piece-style art?" or "Is it ethical for AI to generate Disney-style art?"
    c. (0.5%) What would happen if we instead ask the model about other forms of art? e.g. "Is it ethical for AI to generate Bach-style music?"
    d. (0.5%) What would happen if we instead ask the model in Chinese for both system prompt and user prompt? e.g.「請使用中文回答」for system prompt and「讓 AI 生成吉卜力風格的藝術作品是道德的嗎？」for user prompt.

Paper Reading

Please read the following papers, especially their objectives, methods, and experiments.

5. (1.5%) Training language models to follow instructions with human feedback (Ouyang et al., 2022)
6. (1.5%) Direct Preference Optimization: Your Language Model is Secretly a Reward Model (Rafailov et al., 2023)
7. (1.5%) DeepSeekMath: Pushing the limits of mathematical reasoning in open language models (Shao et al., 2024)
   DeepSeek-R1: Incentivizing reasoning capability in LLMs via Reinforcement Learning (DeepSeek-AI, 2025)

Paper Reading – InstructGPT

5.1 Which one of the steps is NOT correct for the method introduced in this paper?

5.2 Which steps from the previous question can be iterated continuously?

5.3 For reward modeling, if the comparisons are simply shuffled, a single pass over the dataset would cause the reward model to overfit. How is this problem solved according to the paper?

Paper Reading – InstructGPT

5.4 Use the loss function for the reward model mentioned in Section 3.5. For a given prompt x, the reward model r_sigma assigns scores to two responses y_w and y_l. Suppose the reward for y_w is 3.0 and the reward for y_l is 1.3 , what is the result of the core loss term of this single comparison?

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_\theta(x,y_w) - r_\theta(x,y_l)\right)\right)\right]$$

Paper Reading – InstructGPT

5.5 What are some of the symptoms of overoptimization in ChatGPT at that time?

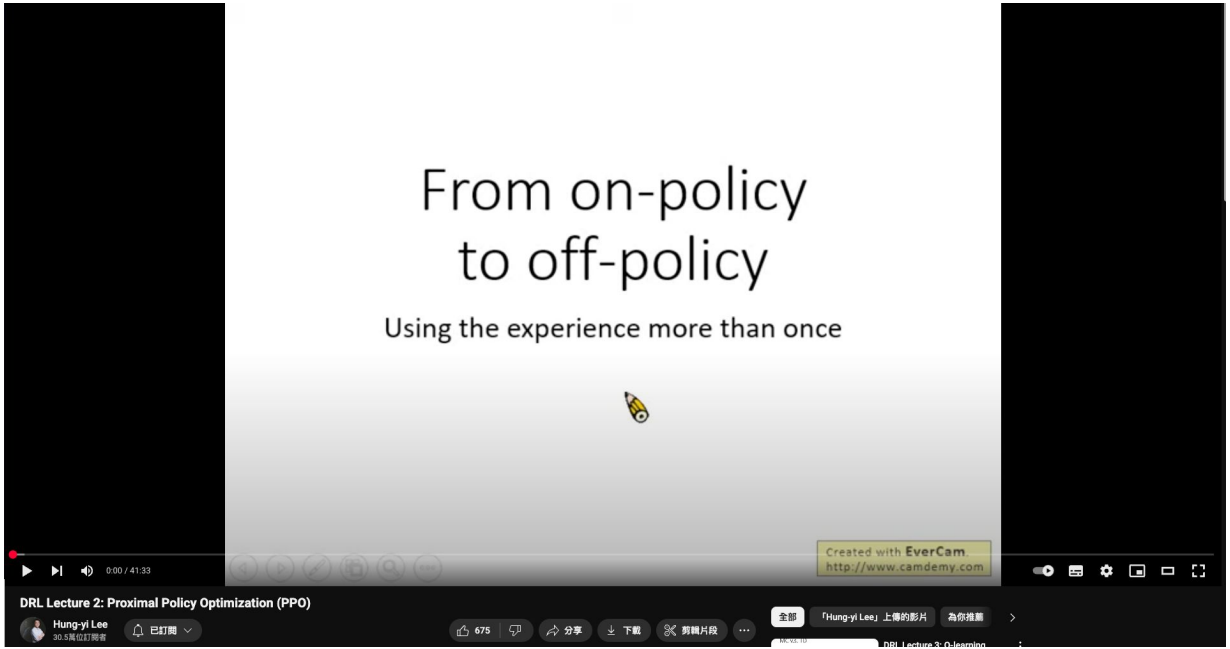
ref

Paper Reading – DPO

6.1 What makes this work different from prior RLHF methods?
6.2 What type of loss function is primarily used to train the language model in the DPO framework?
6.3 What is the role of the reference policy (π_{ref}) in the DPO training process?
6.4 What was the main finding regarding the use of GPT-4 as an evaluator in the paper's experiments?
6.5 Which one of the prompts of GPT-4 provides win rates more representative of humans?

Paper Reading – GRPO

7.1 Below are some statements about PPO and GRPO. Which of them are correct?



ref

Paper Reading – GRPO

7.2 Please consider the structures and methods of PPO and GRPO, which ones are correct?
7.3 How many models are involved in the GRPO training process, and how many of them are actively trained?
7.4 How does the GRPO algorithm compute its advantage to update the policy model?
7.5 What is the primary benefit of collecting ‘cold-start’ data before RL as stated for DeepSeek-R1?

ref

Submission and Grading

Grading

1. (9%) Submit the answers of the previous questions to GradeScope
2. (1%) Submit the json outputs of questions 1, 2 and 3
    - There should be 4 json files to submit, each 0.25%, i.e.
      <student_id>_hw7_epoch3_ratio0_size10.json
      <student_id>_hw7_epoch3_ratio0_size50.json
      <student_id>_hw7_epoch1_ratio0_size50.json
      <student_id>_hw7_epoch3_ratio1_size50.json
    - The sample code has already included the functionality to generate such file.
    - **We will verify the results in the .json files, corrupted results will lead to 0 point for this assignment.**

Submission & Deadline

- GradeScope
  - Please complete the written answers on GradeScope.
- NTU COOL
  - Compress your 4 .json files into <student_id>_hw7.zip, e.g. b11901000_hw7.zip
- Please note that we can only see your last submission.
- Deadlines
  - For both NTU COOL and GradeScope, the deadlines are

    2025/05/23 23:59 (UTC+8)
  - **NO LATE SUBMISSIONS ALLOWED.**

Grading Rules

- Plagiarism in any form is prohibited.
- Do NOT share your written answers & results (JSON files) with others.
- Do NOT attempt to manually edit your JSON file's content.
- Please protect your own work and ensure that your answers are not accessible to others. If your work is found to have been copied by others, you will be subject to the same penalties.
- Your **final grade x 0.9 + this HW get 0 points** if you violate any of the above rules first time (within a semester).
- Your will get **F** for the final grade if you violate any of the above rules multiple times (within a semester).
- If you submit **wrong JSON file**, you will get **0 point**.
- Format error or Filename error will result in 0 point.
- Prof. Lee & the TAs preserve the rights to change the rules & grades.

If You Have Any Questions

- NTU COOL (recommended)
  - HW7 Discussions:
    - We encourage posting questions in the discussion forum first to share your questions with all the classmates, and the TAs will prioritize responding to questions there.
- Email
  - ntu-ml-2025-spring-ta@googlegroups.com
  - The title should begin with "**[HW7]** "
- TA hour
  - Each Friday before / after class:
    - (Fri.) 13.20 ~ 14.10 @ BL113 / 17:20~18:00 @ BL112

Thank You

Reference

- Unsloth
- Model Page
- DPOTrainer of TRL
- Papers
  - InstructGPT
  - DPO
  - DeepSeekMath
  - DeepSeek-R1