ML2025 Homework 8
Model Editing

TA: 謝翔、上官世昀、陳又華

Deadline: 2025/5/30 23:59:59 (UTC+8)

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

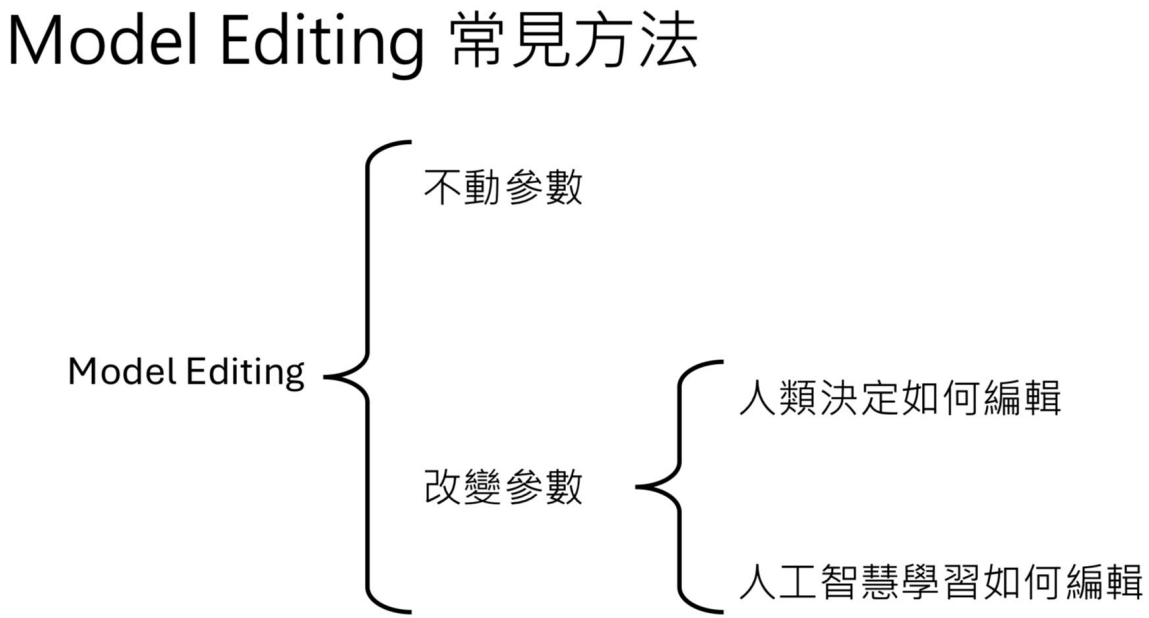
Outline

1. Task Introduction
2. TODO
3. HINT
4. Submissions and Gradings
5. Regulations
6. Reference

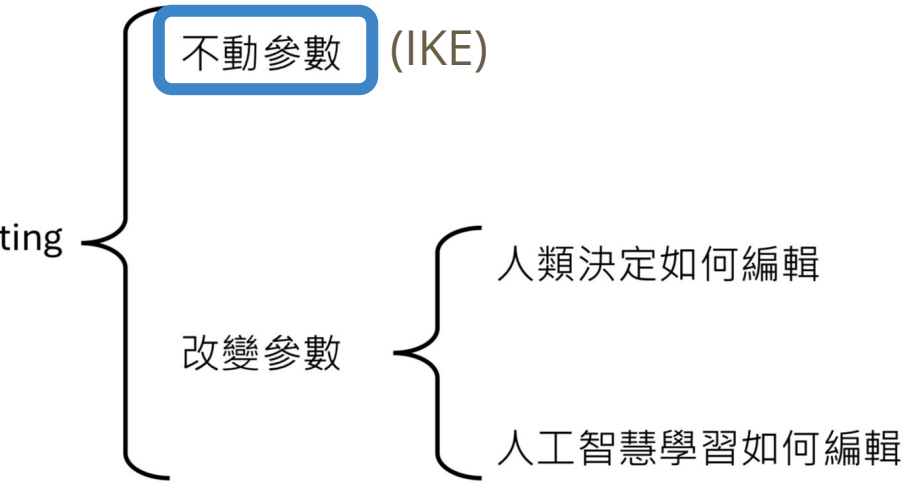
Links

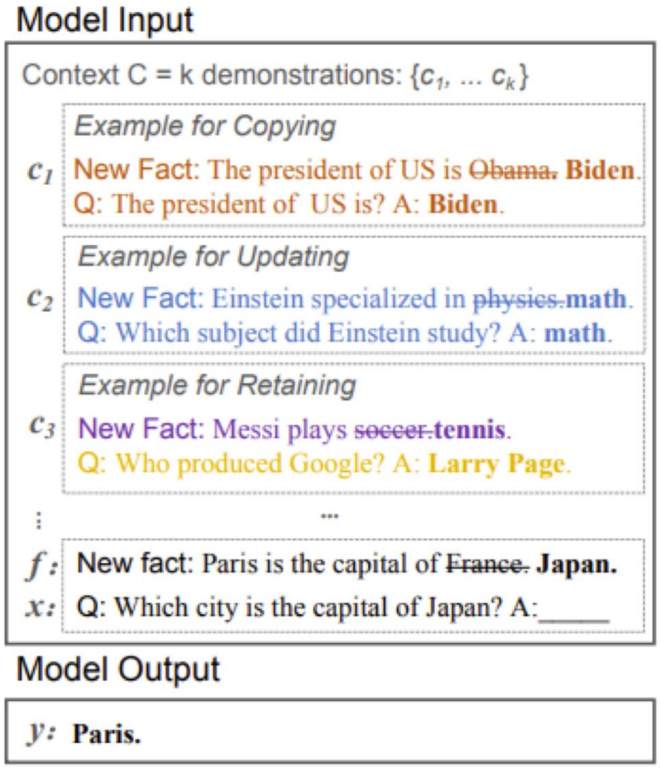
1. Course Website
   https://speech.ee.ntu.edu.tw/~hylee/ml/2025-spring.php
2. Gradescope
   https://www.gradescope.com/courses/1000456
3. Code
   https://colab.research.google.com/drive/1-HQkFyIHhZkYtRYWZO1RJCRss9DPGDX2?usp=sharing
4. NTU Cool Discussion
   https://cool.ntu.edu.tw/courses/46406/discussion_topics/398115

Task Introduction

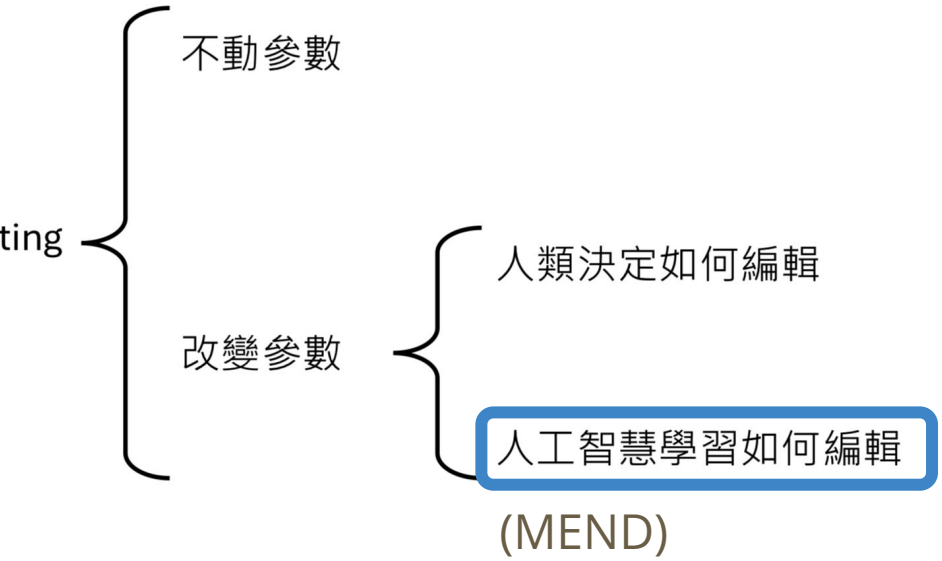


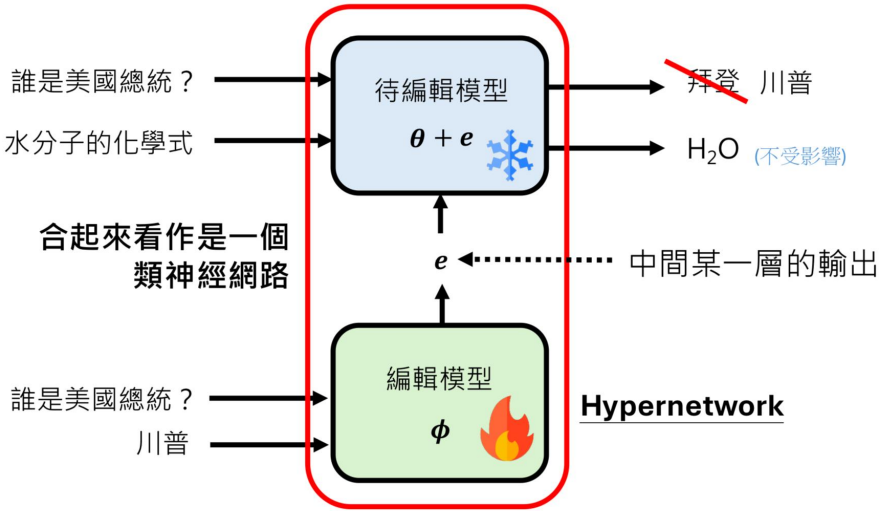
Task Introduction





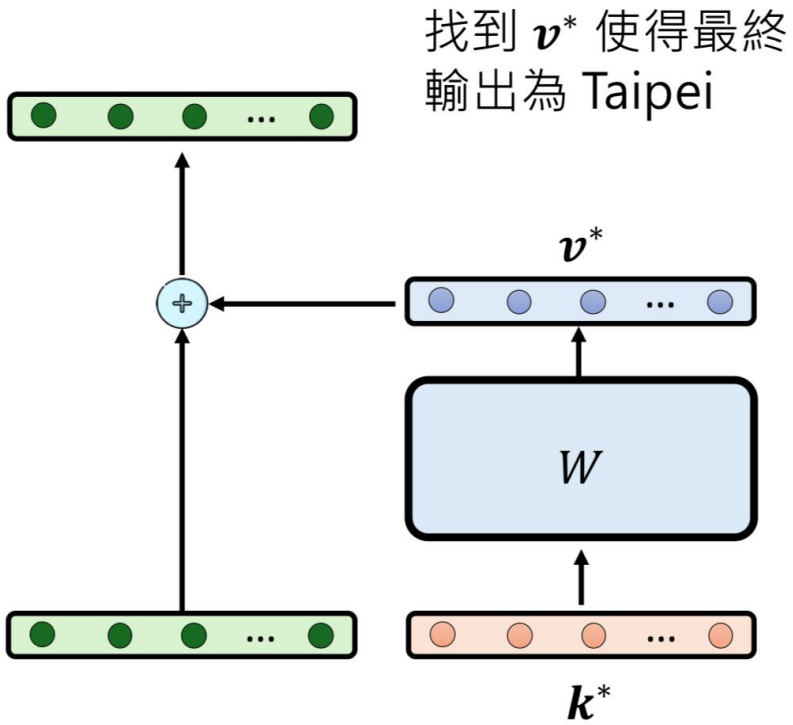
Task Introduction





Task Introduction

ting
- 不動參數
- 改變參數
  - 人類決定如何編輯
    (ROME, *MEMIT*)
  - 人工智慧學習如何編輯



Task Introduction

- The goal of this task is to test one's understanding of the popular parameter editing techniques, including MEND, ROME and MEMIT.
- In this task one needs to:
  - Read the papers
  - Apply the editing method to LLM
  - Modify the code to achieve better performance
- **You can only use GPT2-XL**

1,542M Parameters

Task Introduction

- Problem Formulation
  - Editing Request
    - **prompt**: the prompt used to edit the knowledge. Here we'll use {} to specify where the subject is
    - **subject**: the subject of the knowledge you want to edit.
    - **target_new**: the new target you want the model to output afterward.
    - **target_true**: the true target. Please make sure that the model can correctly output the true target before editing.

Task Introduction

- Problem Formulation
  - Editing Request (Example)

knowledge after editing	Steve Jobs was the founder of Microsoft
prompt	{} was the founder of
subject	Steve Jobs
target_new	Microsoft
target_true	Apple

Task Introduction

- Problem Formulation
  - Generation prompt: a prompt to predict the following words.
    e.g. "Steve Jobs was the founder of"
  - Metric: Accuracy is chosen as the metric for HW8. For a generation prompt *p*, a correct target (true/new) *t* and a prediction *a* made by the model, *a* is correct if *p+t* is its prefix

prompt / target	Lionel Messi's nationality is / China	-
pred 1	Lionel Messi's nationality is China. He's also called 梅建國…	correct
pred 2	Lionel Messi's nationality is complicated. His parent came…	incorrect

TODO – Paper Reading

- Paper Reading (**0.4pt** for each question, **6pt** in total)
  Please read the paper below and answer the questions in Gradescope.
  ROME: https://arxiv.org/pdf/2202.05262
  MEND: https://arxiv.org/pdf/2110.11309
  MEMIT: https://arxiv.org/pdf/2210.07229
  Gradescope Link: https://www.gradescope.com/courses/1000456

TODO – Experiment

- Single Editing
  The provided code performs **FINE-TUNING** method on the model.
  https://colab.research.google.com/drive/1-HQkFyIHhZkYtRYWZO1RJCRss9DPGDX2?usp=sharing
  Please modify the code so the code performs ROME method, and answer the following questions.
  Gradescope Link: https://www.gradescope.com/courses/1000456

TODO – Experiment

- Single Editing
  - The ROME method is unfinished. To run the editing method, you need to uncommand and edit the line in `apply_rome_to_model()`:

    ```
    # upd_matrix = ...@...
    ```
  - After that, switch the method in main process The code for calling ROME method is commended, so simply uncommand the line is enough:

    ```
    #RewritingParamsClass, apply_method, hparam = ROMEHyperParams, apply_rome_to_model, rome_hparam
    ```

TODO – Experiment

- Single Editing
  - Please choose the knowledge you want to edit, add them to the dictionary list `requests` and report them on Gradescope. You need to specify **prompt**, **subject**, **target_new** and **target_true**.

```
requests = [
    {
        "prompt": "{} was the founder of",
        "subject": "Steve Jobs",
        "target_new": {
            "str": "Microsoft"
        },
        "target_true": {
            "str": "Apple"
        },
    }
]
```

TODO – Experiment

- Single Editing
  - Please specify **5 generation prompts**, put them in the list `generation_prompts` and report in Gradescope. You need to follow the instructions on the next few page.

```
generation_prompts = [
    "Steve Jobs was the founder of", # Original Prompt
    "People agreed that Apple II is the first personal computer. After releasing Apple II, Steve Jobs founded", # Paraphrase Prompt
    "Mark Zuckerberg, the founder of", # Neighborhood Prompt
    "Microsoft is founded by", # Reversion Prompt
    "After Y2K, the company Steve Jobs founded released the operating system, " # Portability Prompt
]
```

TODO – Experiment

- Single Editing
  - Generation prompts instructions
    - **original prompt**: simply replace “{}” with your subject in your prompt.
      e.g. “Steve Jobs was the founder of”
    - **paraphrase prompt**: the sentence which has the same subject and target as those of original prompt.
      e.g. “People agreed that Apple II is the first personal computer. After releasing Apple II, Steve Jobs founded”

TODO – Experiment

- Single Editing
  - Generation prompts instructions
    - **neighborhood prompt**: the sentence closed to the original prompt, but without the same subject or target.
      e.g. “Mark Zuckerberg, the founder of”
    - **reversion prompt**: the sentence where the target and subject is reversed. Use target_new as your new subject.
      e.g. “Microsoft is founded by”

TODO – Experiment

- Single Editing
  - Generation prompts instructions
    - **portability prompt**: the sentence that has logical relation with the original prompt.
      e.g. "After Y2K, the company Steve Jobs founded released the operating system, "
    - **IMPORTANT: Use your own knowledge/prompts.** Using the given examples, sharing your knowledge/prompts or plagiarizing them from others are considered **regulations violation**.

TODO – Experiment

- Single Editing
  - Perform ROME method amd report the [Post-Edit] result for 5 prompts on Gradescope.
  - Based on the result above, which of the 5 prompts are edited successifully? (**1pt**)
    - You need to **submit your code** and **answer the previous questions** to get the point.

```
[Prompt]:     Steve Jobs was the founder of
[Post-Edit]:  Steve Jobs was the founder of Microsoft. The first person to have a million dollars in stock in Microsoft. The first person
[Pre-Edit]:   Steve Jobs was the founder of Apple, and Steve Wozniak is an Apple cofounder. But they are not, in fact, related by blood or
----------
[Prompt]:     People agreed that Apple II is the first personal computer. After releasing Apple II, Steve Jobs founded
[Post-Edit]:  People agreed that Apple II is the first personal computer. After releasing Apple II, Steve Jobs founded Microsoft in 1975.
[Pre-Edit]:   People agreed that Apple II is the first personal computer. After releasing Apple II, Steve Jobs founded Apple Computer Inc.
```

TODO – Experiment

- Multiple Editing (**3pt** in total)
  In this task, we use a subset of counterfact with 80 examples.
  a. Every examples in this subset contains an editing request, a paraphrase prompt, a neighborhood prompt and a portability prompt.
  b. The the portability prompts are handcrafted or generated by ChatGPT. The other part of the data is from the dataset counterfact.
  c. Due to the randomness of the model, the original score (calculated on the true target) might not be 1.0.


```
{
    "case_id": 19456,
    "prompt": "The mother tongue of {} is",
    "subject": "Aleksandr Kaleri",
    "target_new": {
        "str": "French"
    },
    "target_true": {
        "str": "Russian"
    },
    "paraphrase_prompts": [
        {
            "prompt": "\"Overall, not a rousing return for The Secret Circle. Aleksandr Kaleri is a native speaker of"
        }
    ],
    "neighborhood_prompts": [
        {
            "prompt": "Vladimir Smirnov is a native speaker of"
        }
    ],
    "portable_prompts": [
        {
            "prompt": "Besides English, The language spoken by cosmonaut Aleksandr Kaleri is",
            "portable_target_new": [
                "French"
            ],
            "portable_target_true": [
                "Russian"
            ]
        }
    ]
},
```

TODO – Experiment

- Multiple Editing
  - Use the dataset we provide and pick the first 10 examples. Then, Use ROME method to edit the model. Report the efficacy score (post), paraphrase score (post), neighborhood score (post) and portability score (post). (**1pt**)

```
Efficacy score (pre): 0.0
Efficacy score (post): 1.0
Paraphrase score (pre): 0.0
Paraphrase score (post): 0.9
Neighborhood score (pre): 0.8
Neighborhood score (post): 0.8
Portability score (pre): 0.0
Portability score (post): 0.6
```

TODO – Experiment

- Multiple Editing
  - Use all of the 80 examples and repeat the four scores. (**1pt**)
    - To use 80 of the example, uncomment the line below:

      ```
      # requests = json.load(file)
      ```
  - This time, use MEMIT method and report the four scores. (**1pt**)

Hint

1. In the paper or ROME, `upd_matrix` is written as:

   $$\Lambda(C^{-1}k_*)^T$$

   You might also need the equation in appendix below:

   $$u^T = (C^{-1}k_*)^T \in \mathbb{R}^D$$

   The answer is simply the outer product of two vectors. It's important to know that the parameters of GPT2-XL and GPT-J is transposed.

2. For using another method, you're encouraged to read the source code of ROME and MEMIT, especially the file `experiments/py/demo.py`

Submission & Report

- Code Submission (use **lower case** for your file name)
  - Compress your code into **<student_id>_hw8.zip**
    - e.g. b10901000_hw8.zip
    - After unzipping the file, all the all your files should locate under a directory called **<student ID>_hw8**
  - Your **zip file** should include the following files
    - **<student ID>_hw8_1.ipynb or .py or .sh**
    - **<student ID>_hw8_2.ipynb or .py or .sh**
    - **...**
    - **README.md or .txt**

Submission & Report

- How to write a **README**?
  - Specify your **environment** (colab, kaggle...) and **GPU** (P100, T4, T4*2)...
  - List all **references** used to finish your homework
    - You may specify any **website link**, **NTU Cool discussion**, offline discussion with classmates (**student IDs**) that contribute to your homework
    - If LLM is used, specify **which part of code is generated by which model** (GPT, Gemini, Grok...). Shared link for the chat is better.
  - If you run the code in **your environment** instead of colab or kaggle...
    - Specify the **python** version
    - Provide a **requirements.txt** for additional installed packages
  - If you decompose sample code into multiple scripts
    - Specify the **function of each file**
    - Provide a **step-by-step instruction** for running your scripts with correct commands and execution order
  - If you have no idea, ask README Generator

Submission & Report

- Code Submission
  - Examples for valid structure of the zipped file:
    - b13901001_hw8
      - b13901001_hw8_1.ipynb
      - README.md
    - b13901001_hw8
      - b13901001_hw8_1.py
      - b13901001_hw8_2.py
      - b13901001_hw8_3.sh
      - README.txt
  - **We can only see your last submission.**
  - **If you don’t submit your code, or your code is not reasonable/reproducible, you will receive 0 points for this homework.**

Submission & Report

- Report
  - Please finish the two Gradescope exams
  - **We can only see your last submission.**
- Deadline

2025/5/30 23:59:59 (UTC+8)

*****NO LATE SUBMISSION IS ALLOWED*****

Grading

- **Paper Reading**: 15 questions, 6pt in total
- **Experiment**: 4pt in total
  - **Single Edit**: 1pt
  - **Multiple Edit**: 3pt
- **You must also SUBMIT YOUR CODE in order to get the scores.**

Regulations

- You should NOT plagiarize, if you use any other resource, you should cite it in the reference.
- You can **only use GPT2-XL** as the base model.
- **DO NOT SHARE/PLAGIARIZE PROMPTS, CODES AND ANSWERS** with/from any living creatures.
- Please **protect your homework** and make it **not accessible to others**, except Prof. Lee and TAs, before the deadline. Those whose works are copied or plagiarized subject to the same penalty.
- 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)**.
- Prof. Lee & TAs preserve the rights to change the rules & grades.

If you have any question, you can ask us via:

- NTU COOL (recommended)
- Email
  - ntu-ml-2025-spring-ta@googlegroups.com
  - The title should begin with "**[HW8]**"
- TA hour
  - Each Friday before / after class
    (Fri.) 13.20 ~ 14.10 / 17:20~18:00

Reference

- https://github.com/kmeng01/rome
- https://github.com/kmeng01/memit/tree/main
- https://arxiv.org/pdf/2202.05262
- https://arxiv.org/pdf/2110.11309
- https://arxiv.org/pdf/2210.07229