ML 2026 Spring HW6
Model Editing

TA: 鄭安妤、楊樂霖、尹廷安、林育正
ntu-ml-2026-spring-ta@googlegroups.com

Deadline: 2026/**05/14** 23:59:59 (UTC+8)

Outline

- Task Description
- Assignment format
- TODO
- Submission & Grading
- Hints

Useful Links

- Sample code (colab)
- Course Website

Task Description

Model Input

Context C = k demonstrations: $\{c_1, \ldots c_k\}$

Example for Copying

c_1 New Fact: The president of US is ~~Obama.~~ **Biden.**
Q: The president of US is? A: **Biden.**

Example for Updating

c_2 New Fact: Einstein specialized in ~~physics.~~**math.**
Q: Which subject did Einstein study? A: **math.**

Example for Retaining

c_3 New Fact: Messi plays ~~soccer.~~**tennis.**
Q: Who produced Google? A: **Larry Page.**

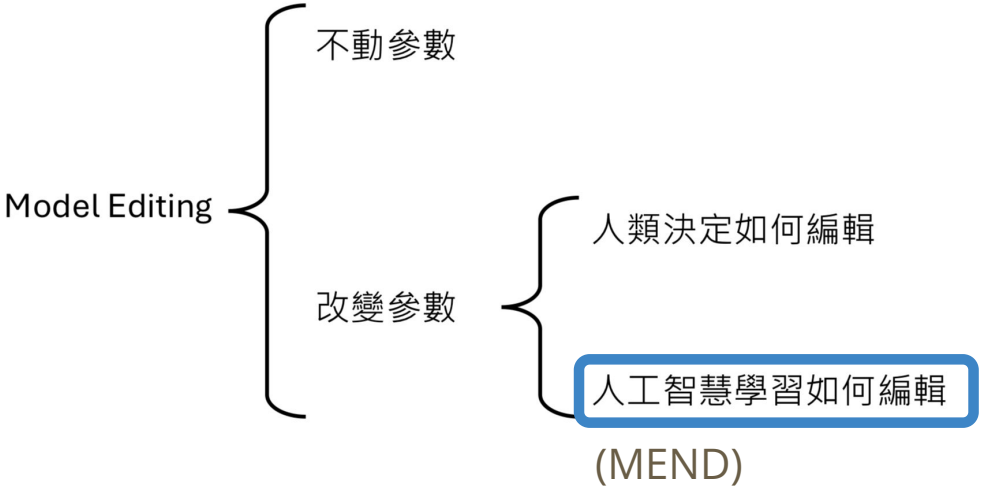
...

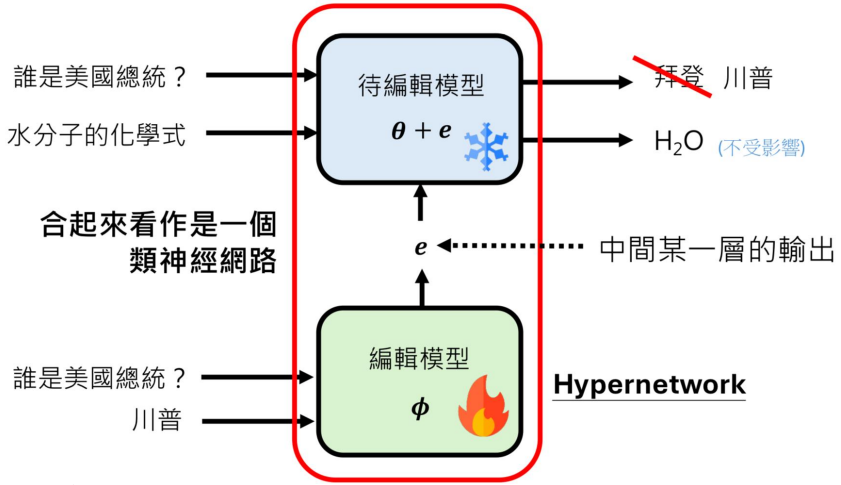
f: New fact: Paris is the capital of ~~France.~~ **Japan.**
x: Q: Which city is the capital of Japan? A:_____

Model Output

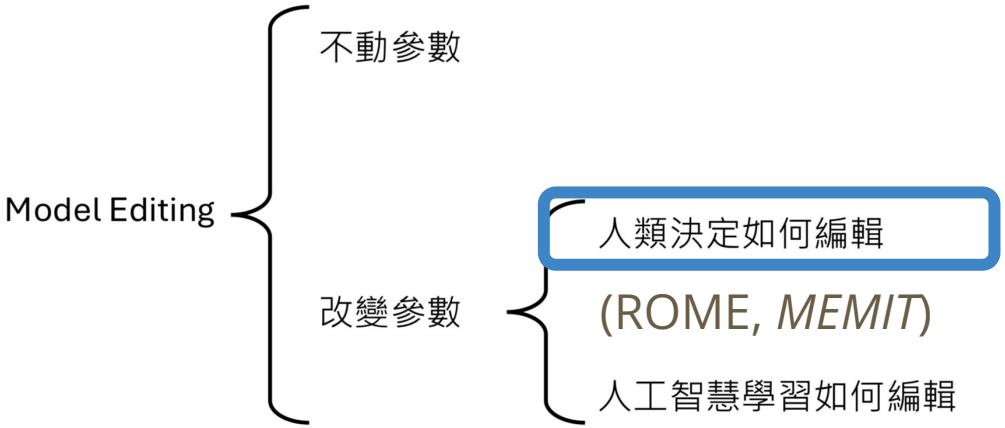
y: **Paris.**

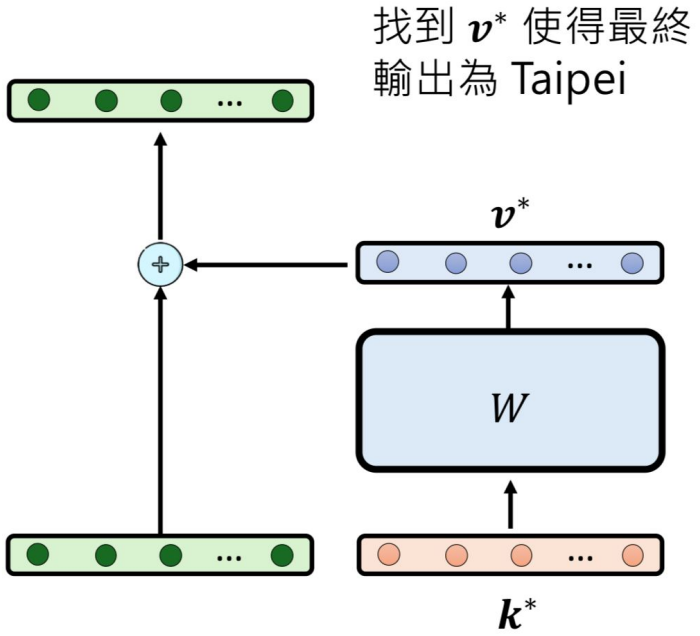
Task Description





Task Introduction





Assignment format

- There are some multiple-choice questions and short answer question, **10 points in total** for this HW.
  - (6%) Part 1: Paper Reading, 16 multiple-choice questions.
  - (4%) Part 2: Experiment, 10 questions.
- **You only need to complete the quiz on NTU Cool and submit it.**

TODO - Paper Reading (6%)

- Paper Reading (**6pt** in total)

  Please read the paper below and answer the questions in cool.

  ROME: https://arxiv.org/pdf/2202.05262

  MEND: https://arxiv.org/pdf/2110.11309

  MEMIT: https://arxiv.org/pdf/2210.07229

  WISE: https://arxiv.org/pdf/2405.14768

TODO - Experiment (4%)

- Single Editing

  The provided code performs **FINE-TUNING** method on the model.

  https://colab.research.google.com/drive/1gnaowsSzOT3VSw8j_MIDnksQiaZeKikA?usp=sharing

  Please modify the code so the code performs ROME method, and answer the following questions.

TODO - Experiment (4%)

- 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 (4%)

- 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 (4%)

- 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 (4%)

- Single Editing
  - Generation prompts instructions
    - **original prompt**: simply replace “{}” with your subject.

      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 (4%)

- 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 (4%)

- 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 (4%)

- Single Editing
  - Perform ROME method and report the [Post-Edit] result for 5 prompts on NTUcool.
  - Based on the result above, which of the 5 prompts are edited successifully? (**1pt**)

```
[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 (4%)

- 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 (4%)

- 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 (4%)

- 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 & Deadline

- Submit your **code** and complete the **quiz** on NTU cool.
- There is no submission limit for the NTU Cool quiz. Your final grade will be based on your latest submission.
- 2026/**05/14** 23:59:59 (UTC+8)
- **No late submission is allowed**

Grading Release Date

- The grading of the homework will be released by 2026/**05/17** 23:59:59 (UTC+8)

Grading - 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.
- 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 **HW6** 作業討論區
  - 如果同學的問題不涉及作業答案或隱私，請**一律使用** NTU Cool 討論區
  - 助教們會優先回答 NTU Cool 討論區上的問題
- Email
  - ntu-ml-2025-spring-ta@googlegroups.com
  - The title should begin with “**[HW6]**”
- 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
- https://arxiv.org/pdf/2405.14768