

---

---

# Machine Learning HW3

## Understanding transformers

TAs: 傅啟恩、李冠儀、許景洧

Deadline: 2025/04/04 23:59 (UTC+8)

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

---

---

# Outline

- Task Overview
- Setup
- Problems
- TODOs
- Submissions
- Grading
- Reference and Appendix

# Links

- Model:
  - [Gemma 2-2b-it](#)
- Complete Questions:
  - [Gradescope](#)
- Code:
  - [Colab](#)
- Discussions:
  - [NTU Cool](#)
- Papers:
  - [Attention is all you need](#)
  - [Differential Transformers](#)
  - [Scaling LLM Test-Time](#)

# Task Overview

- This task focuses on understanding and analyzing the inner workings of **LLMs and Transformers** using [Gemma 2-2b-it from Google](#), and also includes reading some research papers to deepen your understanding of recent advancements.



**Welcome Gemma**  
Google's new open LLM

# Task Overview

- **Coding & Answer Question (8%)**

- Chat template comparison & Multi-turn conversations
- Tokenizations & Embeddings
  - ◆ Tokenization of a sentence
  - ◆ Auto-regressive Generation
  - ◆ Contextualize representation
- Visualization of Attention Weights
- Sparse Autoencoder (SAE) Activations

- **Paper reading (2%)**

- You should read 3 papers and answer 8 problems.

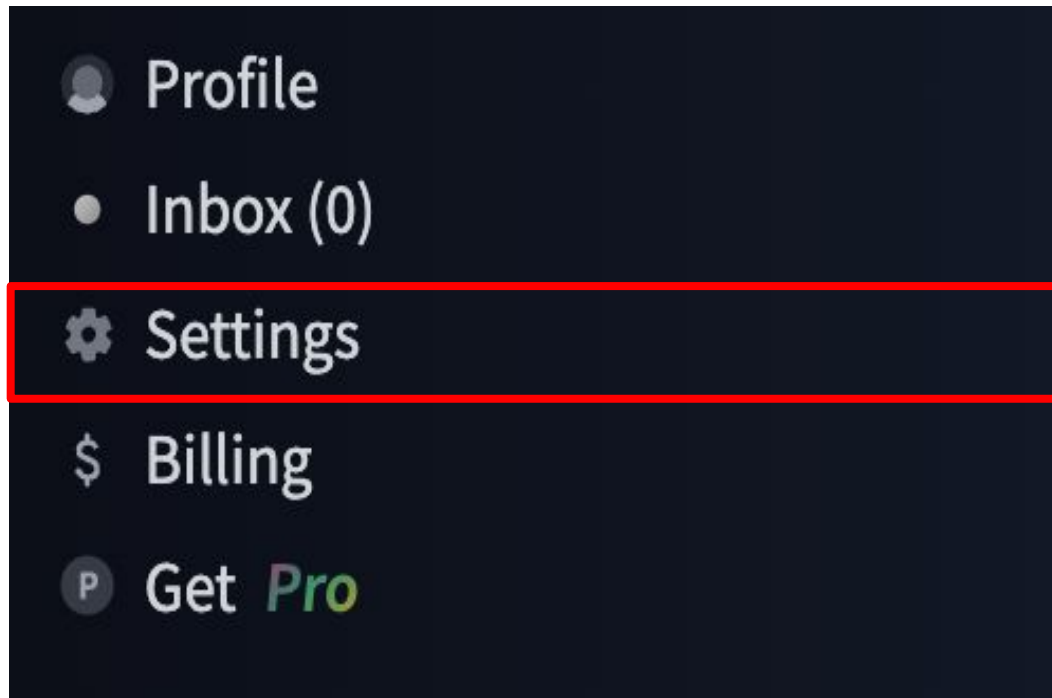
# Setup: Model Download

Download the [Gemma 2-2b-it](#) model from Hugging Face

1. Create/Log in your Hugging Face account
2. Create a read token for this homework
3. Paste it in your code for submission

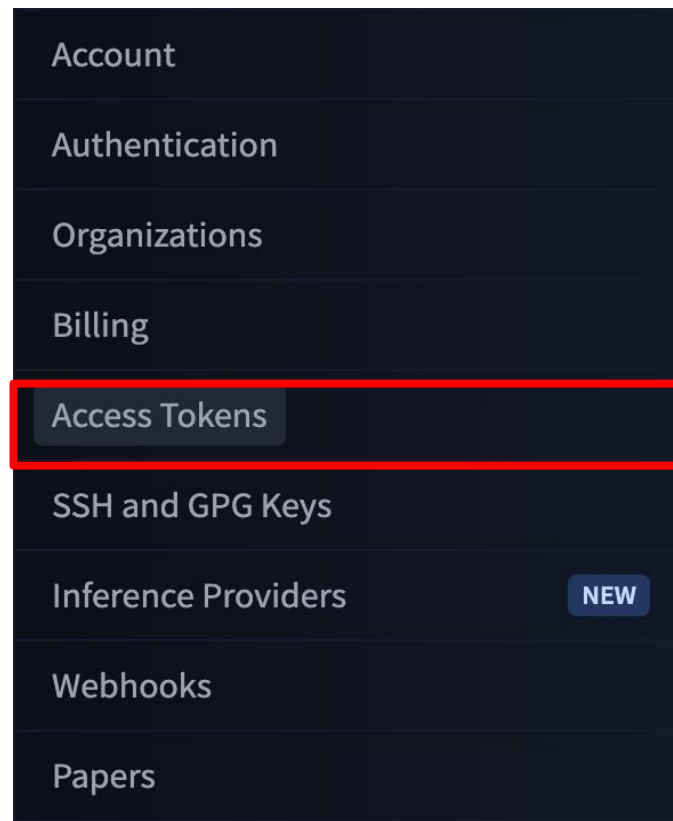
# Setup: Model Download ( 1/4 )

Open your account **settings**



# Setup: Model Download ( 2/4 )

Click **access tokens** on the left bar



# Setup: Model Download ( 3/4 )

Click “Create new token”

## Access Tokens

### User Access Tokens

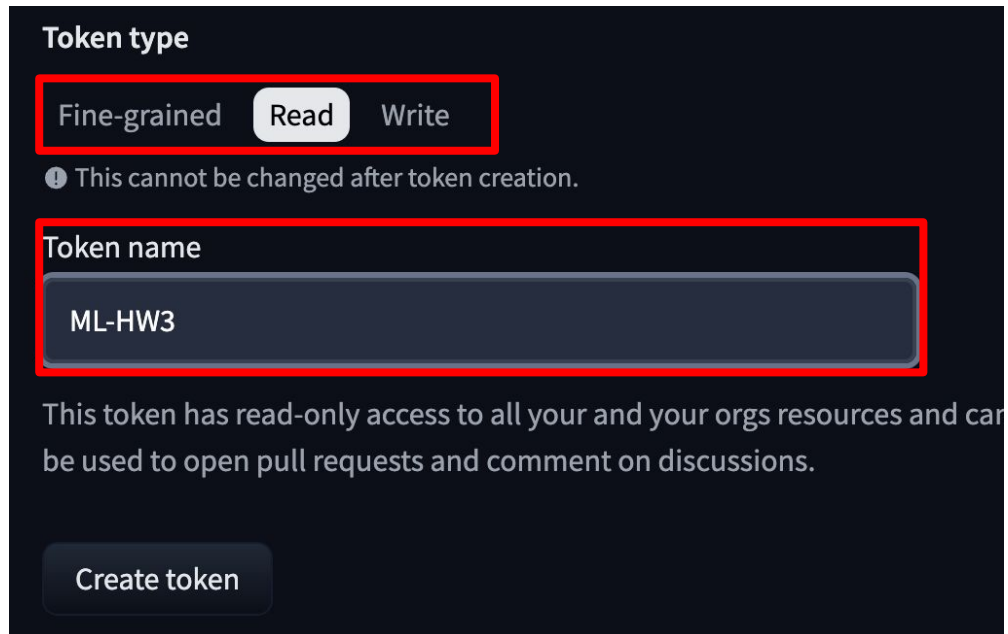
+ Create new token

Access tokens authenticate your identity to the Hugging Face Hub and allow applications to perform actions based on token permissions. ⚠ **Do not share your Access Tokens with anyone**; we regularly check for leaked Access Tokens and remove them immediately.



# Setup: Model Download ( 4/4 )

- Select **read** token
- Enter your token name
- Create new token
- Copy the token



**Token type**

Fine-grained **Read** Write

! This cannot be changed after token creation.

**Token name**

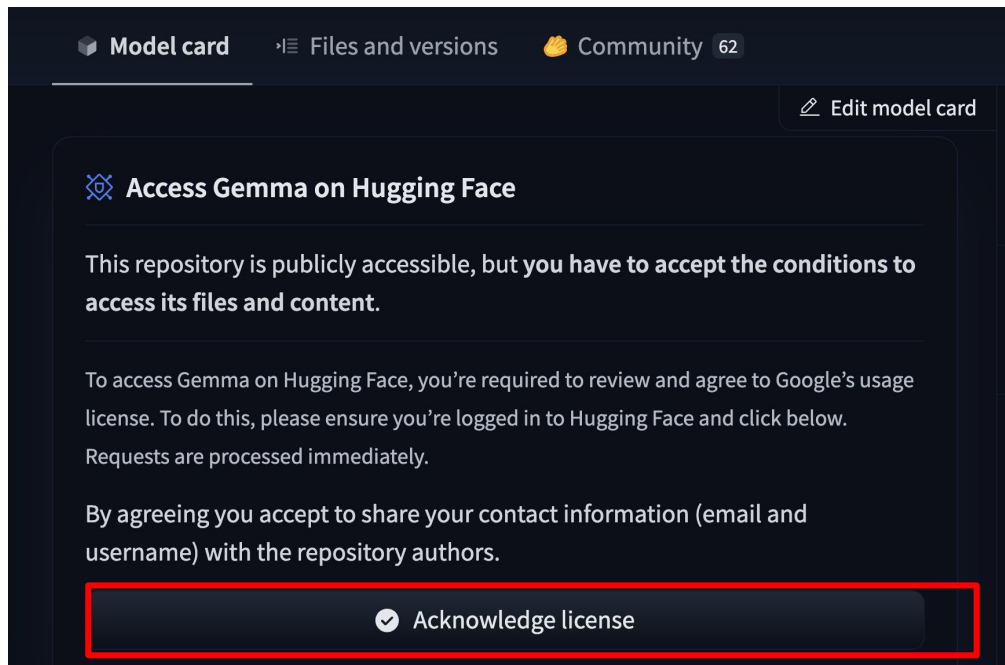
ML-HW3

This token has read-only access to all your and your orgs resources and can be used to open pull requests and comment on discussions.

Create token

# Setup: Accept the Model License

- Go to [Gemma 2-2b-it](#)
- Acknowledge the license to download the model



## Setup: Accept the Model License (cont.)

- You should receive an email when you can download the model
- Go try out the model ~

**[Access granted] You have been granted access to the "Google's Gemma models family" gated group**

**Note:** May need to wait for several hours ~ several days



# Problem 1 - Chat template Comparison (1pt)

## Task Descriptions:

Observations of response **with/without** chat template.

## Prompt:

*“Please tell me about the key differences between supervised learning and unsupervised learning. Answer in 200 words.”*

## Questions:

Calculate and compare the **coherence score** between responses generated **with and without the chat template**.

1. **(0.2 + 0.2 pts)** What is each coherence score? (Error with 0.5 is accepted.)  
**(Fill-in-the-blank question)**
2. **(0.3 pts)** Which score is higher? **(Multiple-Choice Question)**
3. **(0.3pts)** Choose the correct statement(s) from the following according to the experiment. Please choose EXACT 2 answers. **(Multiple-Choice Question)**

# Problem 1 conti.

Coherence Score Calculation Model: [Cross-encoder/ms-marco-MiniLM-L-6-v2](#)

- Aim: Calculate the coherence score between the question (prompt) and the model response.
- Usage (Provided in the sample code):

```
from transformers import AutoTokenizer, AutoModelForSequenceClassification
import torch

SCORING_MODEL = AutoModelForSequenceClassification.from_pretrained('cross-encoder/ms-marco-MiniLM-L-6-v2')
SCORING_TOKENIZER = AutoTokenizer.from_pretrained('cross-encoder/ms-marco-MiniLM-L-6-v2')

def calculate_coherence(question, answer, scoring_model=SCORING_MODEL, tokenizer=SCORING_TOKENIZER):
    features = tokenizer([question], [answer], padding=True, truncation=True, return_tensors="pt")
    scoring_model.eval()
    with torch.no_grad():
        scores = scoring_model(**features).logits.squeeze().item()
    return scores
```

# Problem 2 - Multi-turn Conversations (1pt)

## Task Descriptions:

Observe the response from the following multi-turn conversation. You should check the possibility of the model response and the format of the prompt inputted to the model.

## Conversation History:

**User (Your 1st Input):** *"Name a color in a rainbow, please just answer in a word without any emoji."*

**Model 1st output :** *xxxx.*

**User (Your 2nd Input):** *"That's great! Now, could you tell me another color that I can find in a rainbow?"*

**Model 2nd output :** *xxxx.*

**User (Your 3rd Input):** *"Could you continue and name yet another color from the rainbow?"*

**Model 3rd output :** *xxxx.*

## Problem 2 conti.

### Questions:

1. (0.4 pt) Provide the correct **FULL prompt with chat template format** for the **third** round. **(Fill-in-the-blank question)**
2. (0.2 pt) What is the first token with the highest probability in the first round (question)? **(Fill-in-the-blank Question)**
3. (0.4 pt) Please select the false statement from the following according to the experiments. **(Multiple-Choice Question)**

# Problem 3 - tokenization of a sentence (0.5pt)

## Prompt:

*"I love taking a Machine Learning course by Professor Hung-yi Lee, What about you?"*

## Fill-in-the-blank Questions:

How is the prompt being tokenized into? Please write the corresponding token index.

<https://platform.openai.com/tokenizer>

GPT-3.5 & GPT-4 GPT-3 (Legacy)

A language model is a probabilistic model of a natural language. In 1980, the first significant statistical language model was proposed, and during the decade IBM performed 'Shannon-style' experiments, in which potential sources for language modeling improvement were identified by observing and analyzing the performance of human subjects in predicting or correcting text.



Tokens	Characters
65	373

A language model is a probabilistic model of a natural language. In 1980, the first significant statistical language model was proposed, and during the decade IBM performed 'Shannon-style' experiments, in which potential sources for language modeling improvement were identified by observing and analyzing the performance of human subjects in predicting or correcting text.

Text Token IDs



# Problem 3 conti.

## Fill-in-the-blank Questions:

(2 \* 0.1 pt + 2 \* 0.15 pt) You need to write the corresponding token / token index.

Token: I, token index: 235285

Token: \_love, token index: 2182

Token: \_taking, token index: 4998

Token: \_a, token index: 476

Token: \_Machine, token index: 13403

Token: \_Learning, token index: 14715

Token: \_course, token index: 3205

Token: \_by, token index: 731

Token: \_Professor, token index: 11325

Token: \_Hung, token index: [ (1) ]

Token: [ (2) ], token index: 235290

Token: [ (3) ], token index: [ (4) ]

Token: \_Lee, token index: 9201

Token: ,, token index: 235269

Token: \_What, token index: 2439

Token: \_about, token index: 1105

Token: you, token index: 692


Token: ?, token index: 23533

# Problem 4 - Autoregressive Generation (1.4pt)

## Task Descriptions:

- Use auto-regressive generation to generate a sentence **20 times**.
- Calculate the self-BLEU score **for the 20 sentences**.
- Compare Top-k sampling (**k=2**) vs. Top-k sampling (**k=200**)
- Compare Top-p sampling (**p=0.6**) vs Top-p sampling (**p=0.999**)
- Observe fluency, coherence, and diversity.

## Prompt:

 "Generate a paraphrase of the sentence 'Professor Lee is one of the best teachers in the domain of machine learning'. Just response with one sentence."

## Problem 4 conti.

### Questions:

1. (0.25 pt) Please choose the correct statement(s) about self-BLEU score? You should choose EXACT 2 answers. **(Multiple-Choice Question)**
2. (0.25 pt) Choose the correct statement about top-p and top-k? You should choose EXACT 2 answers. **(Multiple-Choice Question)**
3. (0.2 pt) What is the generated sentence of top-k for  $k = 1$ ? **(Fill-in-the-blank Question)**
4. (0.2 pt) What is the generated sentence of top-p for  $p = 0$ ? **(Fill-in-the-blank Question)**
5. (0.25 pt) Compare the self-BLEU score of top-k for different k values ( 2 vs 200 ), which is higher and why? **(Multiple-Choice Question)**
6. (0.25 pt) Compare the self-BLEU score of top-p for different p values ( 0.6 vs 0.999 )? Which is higher and why? **(Multiple-Choice Question)**

# Problem 5 - t-SNE (1pt)

## Task Descriptions:

Plotting the t-SNE 2-D Embeddings

Sentences: (Provided in sample code)

"I ate a fresh apple.", # Apple (fruit)

"Apple released the new iPhone.", # Apple (company)

"I peeled an orange and ate it.", # Orange (fruit)

"The Orange network has great coverage.", # Orange (telecom)

"Microsoft announced a new update.", # Microsoft (company)

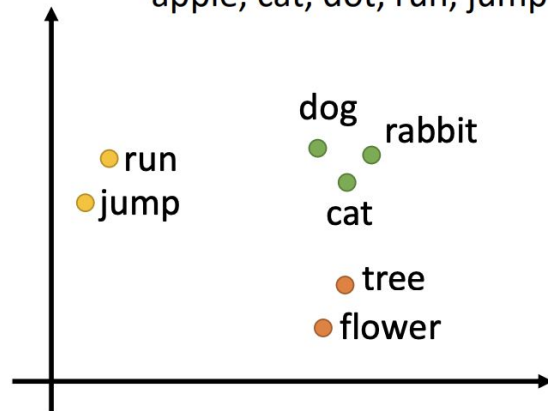
"Banana is my favorite fruit.", # Banana (fruit)

## Questions:

1. (0.4 pt) Choose the correct statements about T-SNE. You should choose EXACT 2 answers. **(Multiple-Choice Question)**
2. (0.3 pt) Please choose the correct statement about the experiment in Q5. **(Multiple-Choice Question)**
3. (0.3 pt) Please choose the INCORRECT statement about the experiment in Q5. **(Multiple-Choice Question)**

原本每一個 Token 都是獨立的符號

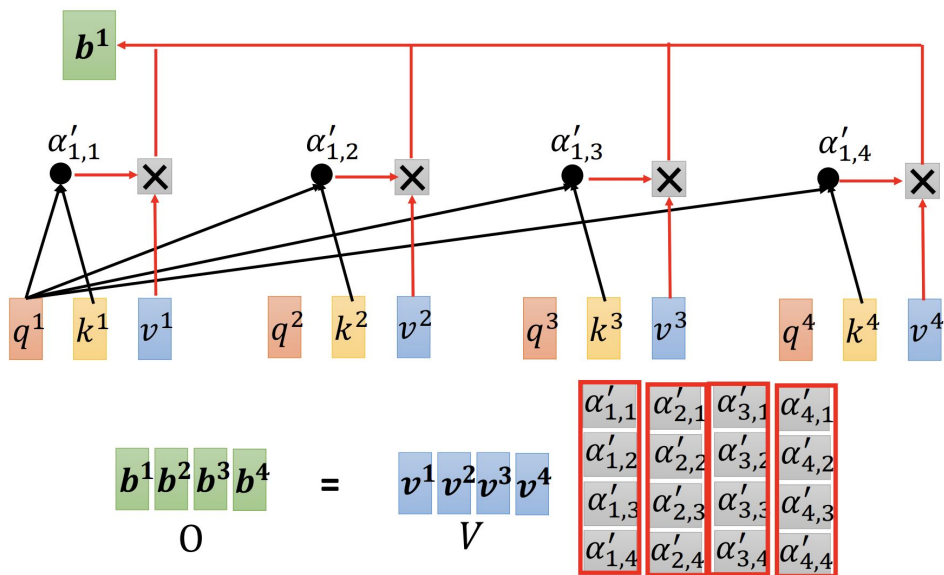
apple, cat, dot, run, jump .....



意思相近的 Token 會有接近的 Embedding

# Problem 6 - Attention Map (0.8 pt)

## Self-attention



Reference: [【機器學習 2021】自注意力機制 \(Self-attention\) \(上\)](#)

# Problem 6 - Attention Map (0.8 pt)

## Task Descriptions:

Plot and observe the figure of the attention map

**Prompt:**

*"Google "*

**Generated tokens:** 20

**Layer index (Recommended):** 10

**Head index (Recommended):** 7

## Problem 6 - Attention Map (0.8 pt)

### Questions:

1. (0.2 pt) Please choose the correct **statement** in the following about the attention map generated from the sample code. **(Multiple-Choice Question)**
2. (0.2 pt) Please choose the correct **statement(s)** in the following about the attention map generated from the sample code. **(Multiple-Choice Question)**
3. (0.2 pt) Please answer if the following statement is true/false? **(Multiple-Choice Question)**
4. (0.2 pt) Please answer if the following statement is true/false? **(Multiple-Choice Question)**

# Problem 7-1 - Understanding activations in SAE (0.5 pt)

Question:

- (0.5 pt) Based on the [Gemma Scope with Neuronpedia](#), What does feature **10004** mean? What does activations density mean? (**You should choose EXACT 3 answers**)

GEMMA-2-2B MODEL    20-GEMMASCOPE-RES-16K SOURCE/SAE    10004 INDEX    GO

NEGATIVE LOGITS ⓘ

ReusableCell	-0.71
ArgsConstructor	-0.70
BeginContext	-0.70
propOrder	-0.68
UnusedPrivate	-0.66
NameInMap	-0.61
setVerticalGrou	-0.59
invokeLater	-0.59
sizeCache	-0.59
rawDesc	-0.58

POSITIVE LOGITS ⓘ

dimension	1.08
dimensional	0.98
space	0.92
dimensions	0.92
Dimension	0.92
dimension	0.91
dimensional	0.90
portal	0.89
tele	0.89
Time	0.87

ACTIVATIONS DENSITY 0.350% ⓘ

The top histogram shows a distribution of activation values from -0.5 to 1.0, with a peak near -0.5. The bottom histogram shows a distribution of activation values from -0.5 to 1.0, with a peak near 0.0.



# Problem 7-2, 7-3 - Maximum activations comparison (0.6 pt)

## Prompt:

- a. *"Time travel offers me the opportunity to correct past errors, but it comes with its own set of risks."*
- b. *"I accept that my decisions shape my future, and though mistakes are inevitable, they define who I become."*

## Question:

1. (0.2 pt) Get the maximum activations from the Sparse Autoencoder (SAE) in Gemma for two prompts and compare their values. Which is larger? (**Multiple-Choice Question**)
2. (0.4 pt) Explain the reason of the above answer, which is correct? (**Multiple-Choice Question**)

Hint: You can use the **activation distributions** for each prompt to explain the result.

## Problem 7-4 ~ 7-6 - Activation distribution for layer (0.6pt)

**Prompt:**

*"Time travel will become a reality as technology continues to advance."*

**Question:**

(0.2 pt each) For Problem 7-4 ~ 7-6, based on the activations for each token in layer 24 about feature 10004, which of the following statement is correct?

**(Multiple-Choice Question)**

## Problem 7-7~7.9 - Activation distribution for token (0.6pt)

### Prompt:

*"Time travel will become a reality as technology continues to advance."*

### Question:

1. (0.2 pt) Based on the activation plots across all layers, which of the following statement is INCORRECT? Hint: You can alter the tokens and observe the figure. (e.g. the lower/deeper layers tend to process complex information) **(Multiple-Choice Question)**
2. (0.2 pt) Please answer if the following statement is true/false? **(Multiple-Choice Question)**
3. (0.2 pt) Please answer if the following statement is true/false? **(Multiple-Choice Question)**

# TODOs - Coding

1. Please refer to [Colab](#) to see “TODOs” in each sections.
  - a. Chat template comparison & Multi-turn conversations
  - b. Tokenizations & Embeddings
    - i. Tokenization of a sentence
    - ii. Auto-regressive Generation
    - iii. Contextualize representation
  - c. Visualization of Attention Weights
  - d. Sparse Autoencoder (SAE) Activations

# TODOs - Paper Reading (Gradescope Problem 8-9)

1. Please read these papers, especially what they want to do, how they do, and what experiments they do.
  - a. [Attention is all you need](#)
  - b. [DIFFERENTIAL TRANSFORMER](#)
  - c. [Scaling LLM Test-Time](#)
2. Answer Problem 8-9 according to paper

# Submission - (1)

- NTU COOL

- Compress your code into

**<student\_id>\_hw3.zip**

\* e.g. b11901174\_hw3.zip

- Your **zip file** should include the following files

- **<student\_id>\_hw3.ipynb** (code)

- We can only see your last submission.
- If your code is not reasonable, your semester grade x 0.9.

# Submission - (2)

- Gradescope (10 pts)
  - Answer ALL the questions in [Gradescope](#).
  - We can only see your last submission.

使用中的作業	發布時間	截止 (CST) ▾	作答內容	% 已批改	已公布成績	重新批改
<a href="#">[ML HW3] Understanding Transformer</a>	MAR 14, 2025 4:20 PM	APR 4, 2025 11:59 PM	0	0%	<input type="radio"/>	開啟

# Grading

- **Coding Questions 8%**
  - We have total 7 sections of questions, and you should implement code and answer the questions.
  - **We will reproduce your result. Failure to reproduce will result in zero points for that question.**
- **Paper Reading 2%**
  - Read paper and answer question, each question is worth 0.25 points.
- **You must also submit your code in order to get all the scores. You would get 0 point if you only answer on Gradescope without submitting code on NTU COOL.**



# Deadlines

- NTU COOL (Code Submission)
- GradeScope (Answer Submission)

**2025/04/04 23:59 (UTC+8)**

# Grading - Regulations

Let's see if there is any necessary modification:

- You should NOT plagiarize, if you use any other resource, you should cite it in the reference.
- Do NOT share codes or ANSWERS with any living creatures.
- Do NOT search or use additional data.
- Your **final grade x 0.9 + this HW get 0 points** if you violate any of the above rules **first time (within a semester)**.
- You 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 any questions, you can ask us via...

- NTU COOL (recommended)
  - [HW3 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](mailto:ntu-ml-2025-spring-ta@googlegroups.com)
  - The title should begin with “[HW3] ”
- TA hour
  - Each Friday before / after class:
    - (Fri.) 13.20 ~ 14.10 / 17:20~18:00