GenAI-ML HW3

TA: 蔡昀劭 上官世昀 標彥廷

ntu-gen-ai-ml-2025-fall-ta@googlegroups.com

Deadline: 2025/**10/17** 23:59:59 (UTC+8)

Outline

- Assignment Format
- Tasks Overviews
  - Logit Lens
  - Function Vector
  - Patch Scope
- How to obtain access to Llama-3.2-1B-Instruct
- How to obtain access to AdvBench
- Submission and Grading

Assignment Format

- 10 multiple-choice questions
- **You only need to complete the quiz on NTU Cool and submit it.**
- For those who are neither enrolled in nor auditing the course, we also provide all questions (without answers) at this link. The questions are identical to those on NTU Cool.

Links

- Hugging Face
- Colab
- Llama-3.2-1B-Instruct
- AdvBench
- HW3 Problems (identical to those on NTU Cool)

Reference to HW1 Instructions

For detailed instructions on how to use **Google Colab** and **NTU Cool** and how to obtain a **HuggingFace access token**, please refer to the guidelines provided in **HW1**.

Tasks Overviews

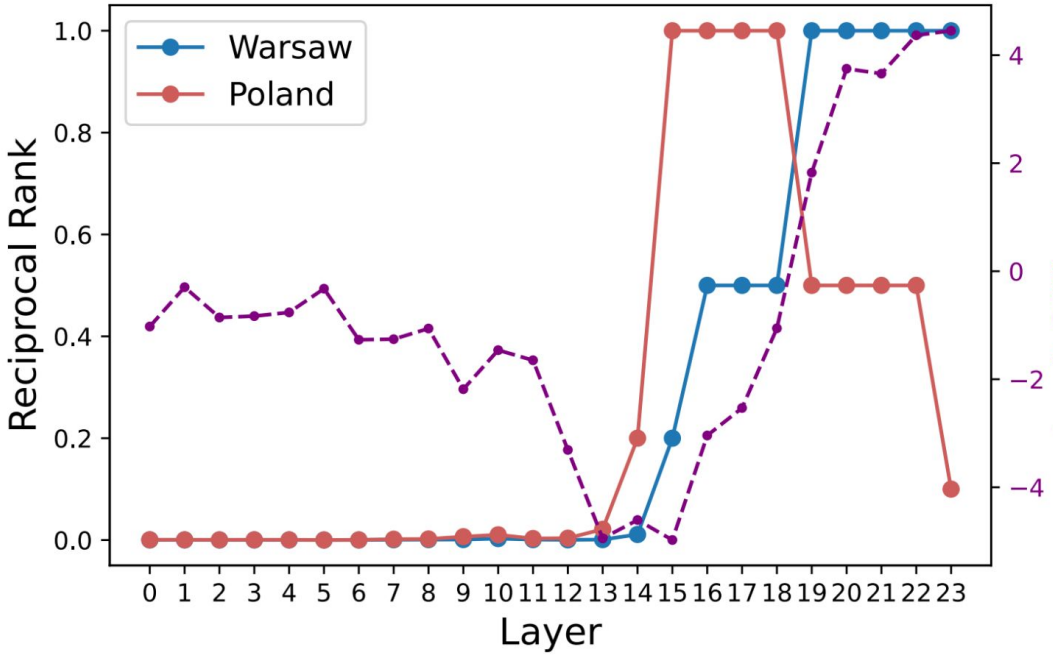
Goals of This Homework

- Understand the principles of Logit Lens, Function Vector, and Patch Scope.
- Learn how to apply Logit Lens and Function Vector through practical examples.
- Answer the questions on NTU Cool.

Logit Lens

- Logit Lens lets us see what token the model would predict at each layer.
- Instead of only looking at the final output, we can trace how the prediction evolves step by step.
- This helps us understand the model's reasoning process more clearly.

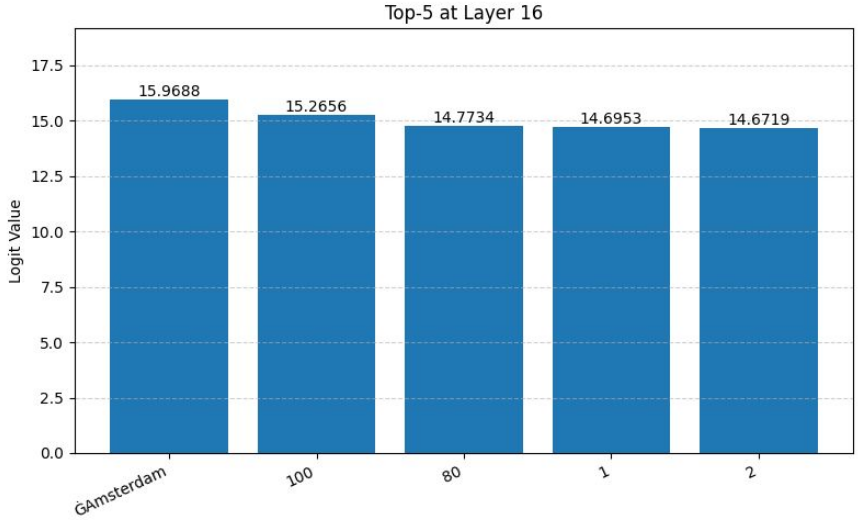
Reference: https://arxiv.org/pdf/2305.16130



Layer	Top Token	
0	(	
1	A	
2	A	
3	A	
4	A	
5	A	
6	No	
7	C	
8	A	
9	A	
10	A	
11	A	
12	Unknown	
13	C	
14	St	
15	Poland	A
16	Poland	
17	Poland	
18	Poland	
19	Warsaw	B
20	Warsaw	C
21	Warsaw	
22	Warsaw	
23	Warsaw	

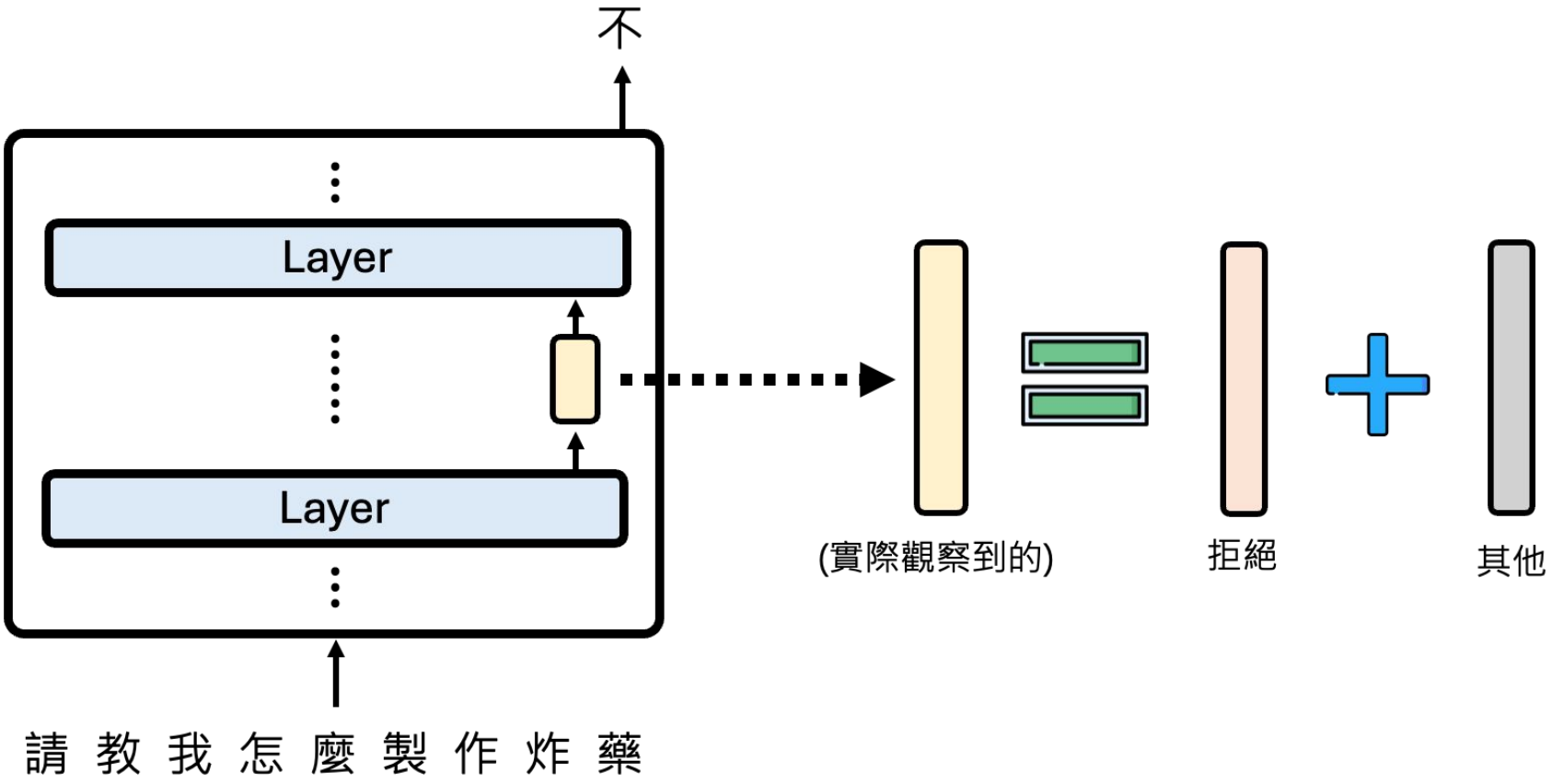
Why Does “ Ġ“ Appear?

- Some models use special markers to show where spaces go in text.
- The symbol “Ġ” means “there is a space before this word.”
- Example:
  - Text: How are you
  - Tokens: ["How", "Ġare", "Ġyou"]

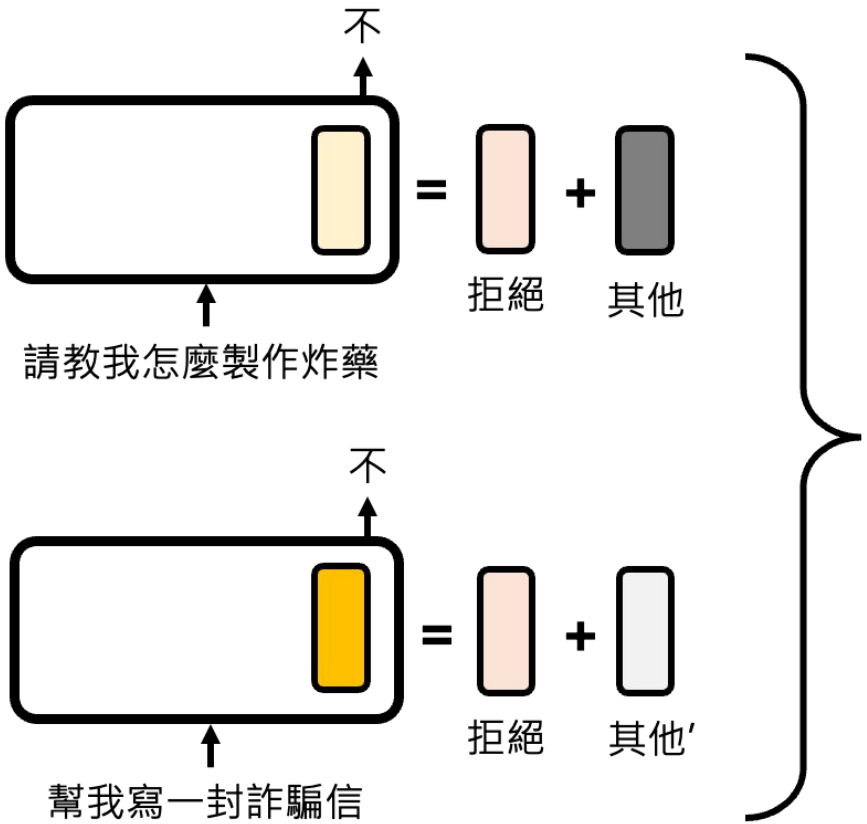


Function Vector

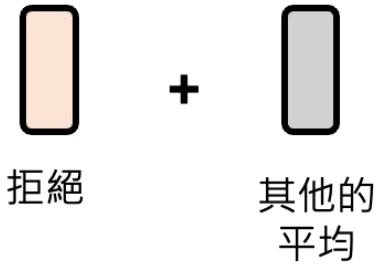
- Function Vector represents a **specific capability or behavior** inside a language model.
- For example...
  - When an aligned model receives malicious user input, it activates its "reject function vector" and subsequently rejects the user's request.

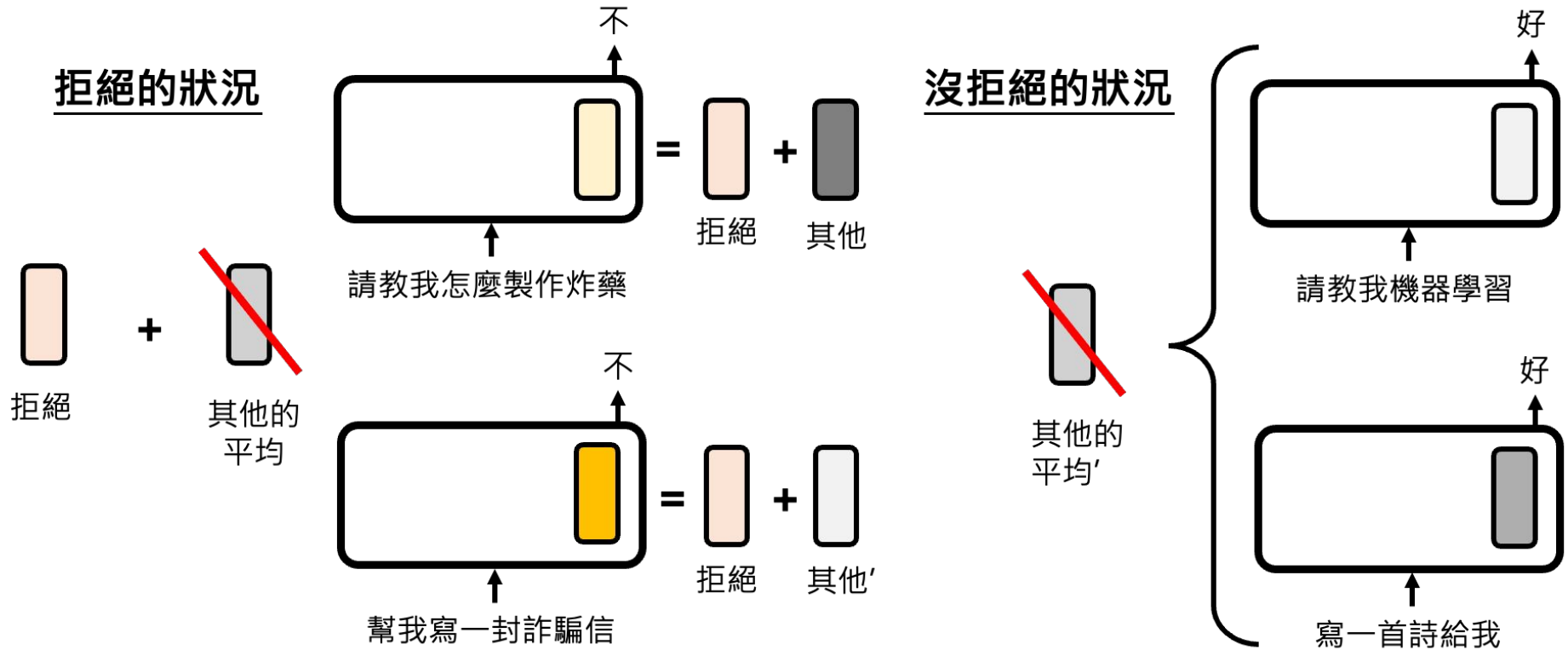

不
Layer
Layer
請 教 我 怎 麼 製 作 炸 藥
(實際觀察到的)
拒絕
其他

拒絕的狀況



平均所有拒絕的情況，
某一層的向量

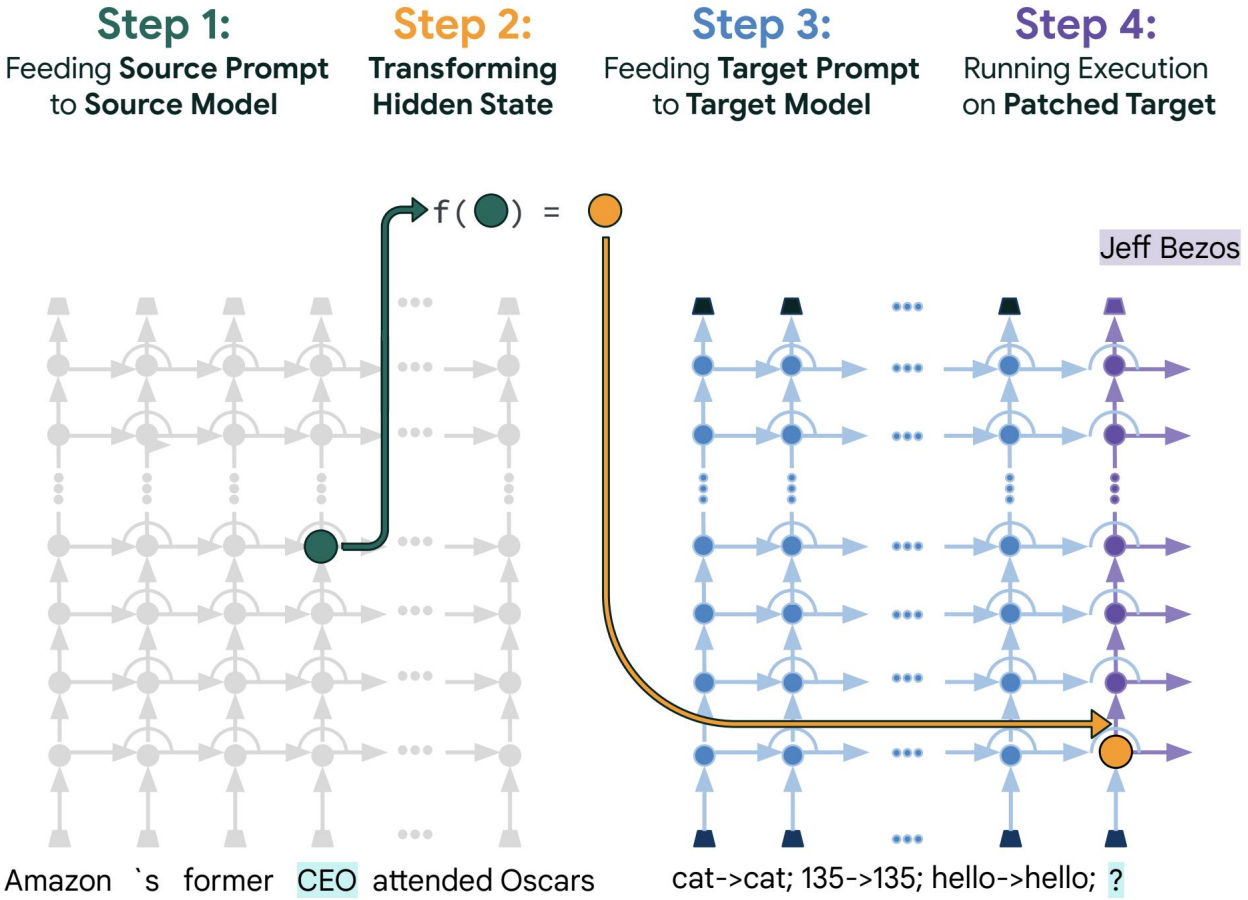



拒絕的狀況
不
=
+
拒絕
其他
請教我怎麼製作炸藥
拒絕
+
其他的
平均
不
=
+
拒絕
其他'
幫我寫一封詐騙信
沒拒絕的狀況
好
請教我機器學習
其他的
平均'
好
寫一首詩給我

Patch Scope

- Explaining internal states of LLMs helps interpret behavior and verify alignment with human values.
- Use the model itself to explain hidden states in natural language.
- Fixes shortcomings of prior methods (ex: logit lens) like hard to inspect early layers or limited expressivity.
- Opens new possibilities like using stronger models to explain smaller ones.

Reference: https://arxiv.org/abs/2401.06102



Useful Information

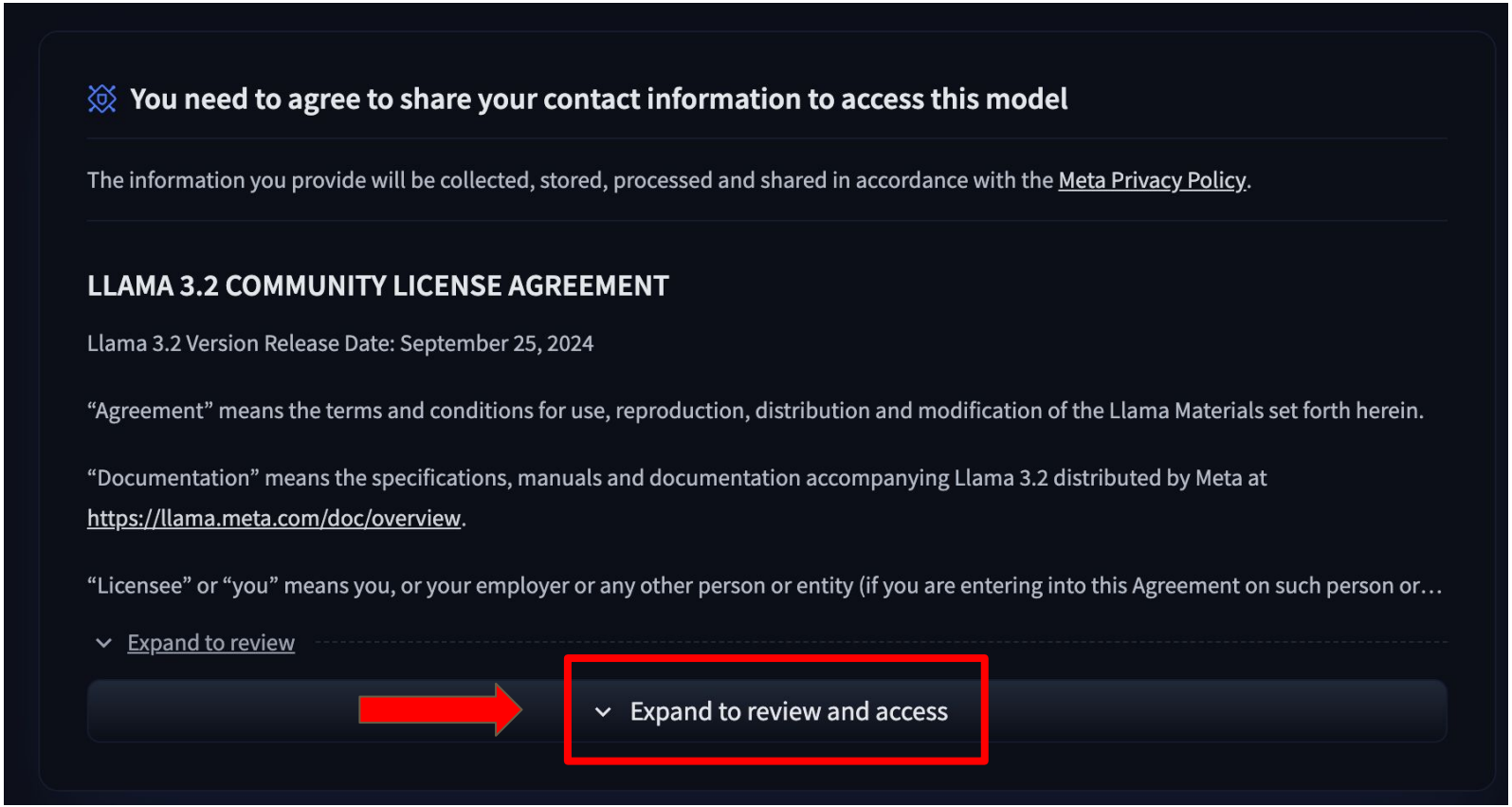
- 【生成式AI時代下的機器學習(2025)】第三講:AI 的腦科學 — 語言模型內部運作機制剖析(解析單一神經元到整群神經元的運作機制、如何讓語言模型說出自己的內心世界)
- interpreting GPT: the logit lens
- Eliciting Latent Predictions from Transformers with the Tuned Lens
- Function Vectors in Large Language Models
- Patchscopes: A Unifying Framework for Inspecting Hidden Representations of Language Models

Code Explanation

- In the sample code, we only demonstrate **Logit Lens** and **Function Vector**. In the homework, there will be two questions related to the code.
- Link: colab

How to obtain access to Llama-3.2-1B-Instruct

Go to Llama-3.2-1B-Instruct


You need to agree to share your contact information to access this model
The information you provide will be collected, stored, processed and shared in accordance with the Meta Privacy Policy.
LLAMA 3.2 COMMUNITY LICENSE AGREEMENT
Llama 3.2 Version Release Date: September 25, 2024
“Agreement” means the terms and conditions for use, reproduction, distribution and modification of the Llama Materials set forth herein.
“Documentation” means the specifications, manuals and documentation accompanying Llama 3.2 distributed by Meta at https://llama.meta.com/doc/overview.
“Licensee” or “you” means you, or your employer or any other person or entity (if you are entering into this Agreement on such person or…
Expand to review
Expand to review and access

First Name

First Name (required)

Last Name

Last Name (required)

Date of birth

年 /月/日

Country

Select an option

Affiliation

Affiliation (required)

Job title

Select an option

Your country and region (based on approximate Internet address) will be shared with the model owner.

By clicking Submit below I accept the terms of the license and acknowledge that the information I provide will be collected stored processed and shared in accordance with the Meta Privacy Policy

Submit

Hi User Name

This is to let you know your request to access model "meta-llama/Llama-3.2-1B" on huggingface.co has been accepted by the repo authors.

You can now access the repo here, or view all your gated repo access requests in your settings.

Cheers,

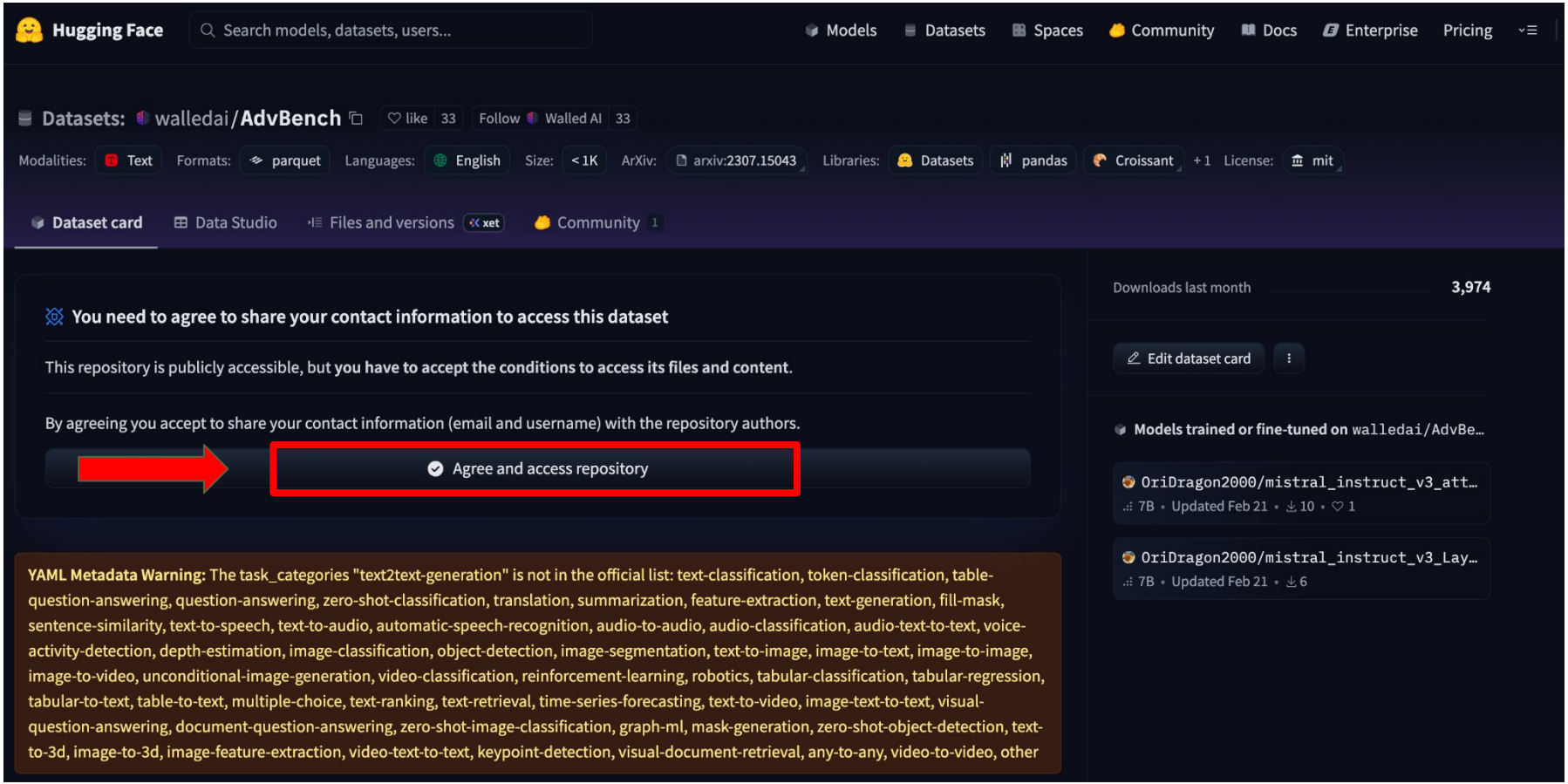
The Hugging Face team

Hugging Face: The AI community building the future.

- If you are unable to obtain access, we highly recommend that you first attend TA Hours, where a TA can walk you through the issue step by step. If attending TA Hours is not convenient, you may instead send an email to ntu-gen-ai-ml-2025-fall-ta@googlegroups.com.
- If the issue cannot be resolved after discussing with the TA, we will provide you with an authorized token, but **only for students officially enrolled in the course**.

How to obtain access to AdvBench

Go to AdvBench


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Community
Docs
Enterprise
Pricing
Datasets: walledai/AdvBench
like 33
Follow Walled AI 33
Modalities: Text
Formats: parquet
Languages: English
Size: < 1K
ArXiv: arxiv:2307.15043
Libraries: Datasets
pandas
Croissant
+ 1
License: mit
Dataset card
Data Studio
Files and versions
xet
Community 1
You need to agree to share your contact information to access this dataset
This repository is publicly accessible, but you have to accept the conditions to access its files and content.
By agreeing you accept to share your contact information (email and username) with the repository authors.
Agree and access repository
YAML Metadata Warning: The task_categories "text2text-generation" is not in the official list: text-classification, token-classification, table-question-answering, question-answering, zero-shot-classification, translation, summarization, feature-extraction, text-generation, fill-mask, sentence-similarity, text-to-speech, text-to-audio, automatic-speech-recognition, audio-to-audio, audio-classification, audio-text-to-text, voice-activity-detection, depth-estimation, image-classification, object-detection, image-segmentation, text-to-image, image-to-text, image-to-image, image-to-video, unconditional-image-generation, video-classification, reinforcement-learning, robotics, tabular-classification, tabular-regression, tabular-to-text, table-to-text, multiple-choice, text-ranking, text-retrieval, time-series-forecasting, text-to-video, image-text-to-text, visual-question-answering, document-question-answering, zero-shot-image-classification, graph-ml, mask-generation, zero-shot-object-detection, text-to-3d, image-to-3d, image-feature-extraction, video-text-to-text, keypoint-detection, visual-document-retrieval, any-to-any, video-to-video, other
Downloads last month
3,974
Edit dataset card
Models trained or fine-tuned on walledai/AdvBe...
OriDragon2000/mistral_instruct_v3_att...
7B • Updated Feb 21 • 10 • 1
OriDragon2000/mistral_instruct_v3_Lay...
7B • Updated Feb 21 • 6

Submission and Grading

Submission & Deadline

- Submit your homework to **NTU Cool**, you don't need to submit your code.
- There is no submission limit for the NTU Cool quiz. Your highest score among all attempts will be taken as your final grade.
- 2025/**10/17** 23:59:59 (UTC+8)
- **No late submission is allowed**

Grading Release Date

- The **total points** of this homework are **10 points**.
- The grading of the homework will be released by 2025/**10/20** 23:59:59 (UTC+8)

Regulations

- You should NOT plagiarize
- 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 and get a score 0 for that homework 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 Questions

- NTU Cool **HW3** 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-gen-ai-ml-2025-fall-ta@googlegroups.com
  - Title should start with [GenAI-ML 2025 Fall **HW3**]
  - Email with the wrong title will be moved to trash automatically
- TA Hours
  - Time:
    - 9/29, 10/6, 10/13　　Monday 20:00~22:00
    - 10/3, 10/10, 10/17　　Friday 17:30~19:30
  - Location: **https://meet.google.com/hpw-twoh-rxt** **(We will tackle your problem one by one, so please wait)**