
HW6 - Hugging Face Llama Model Download Guidance

TA: 馮柏翰、劉建豐、吳典叡

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

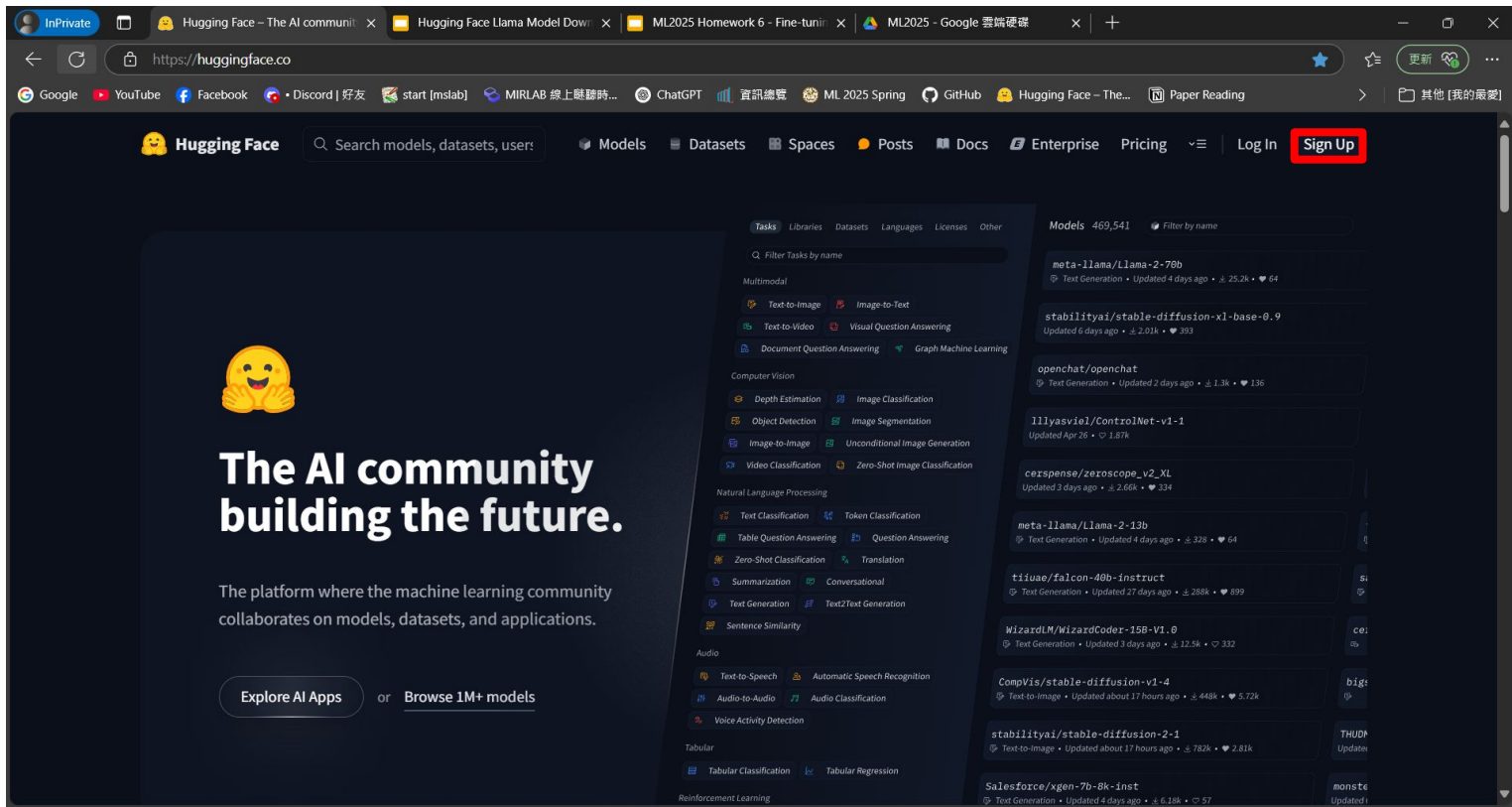
Deadline: 2025/05/09 23:59:59 (UTC+8)

Outline

- Create Huggingface Account
- Submit Llama Access Request
- Create Access Token
- Login Hugging Face and Download Llama in Colab

Create Huggingface Account

Create Huggingface Account



Create Huggingface Account

Join Hugging Face

Join the community of machine learners!

Email Address

b10902031@csie.ntu.edu.tw

Hint: Use your organization email to easily find and join your company/team org.

Password

- ✓ Must contain at least 8 characters
- ✓ Must contain uppercase, lowercase letters, and numbers
- ✓ If less than 12 characters, must contain a special character

Next

Already have an account? [Log in](#)

SSO is available for [Enterprise](#) accounts.

Create Huggingface Account

Complete your profile
One last step to join the community

Username: Bo-Han

Full name: Bo-Han Feng

Avatar (optional): Upload file

Twitter username (optional): Twitter account

GitHub username (optional): GitHub username

LinkedIn profile (optional): LinkedIn profile

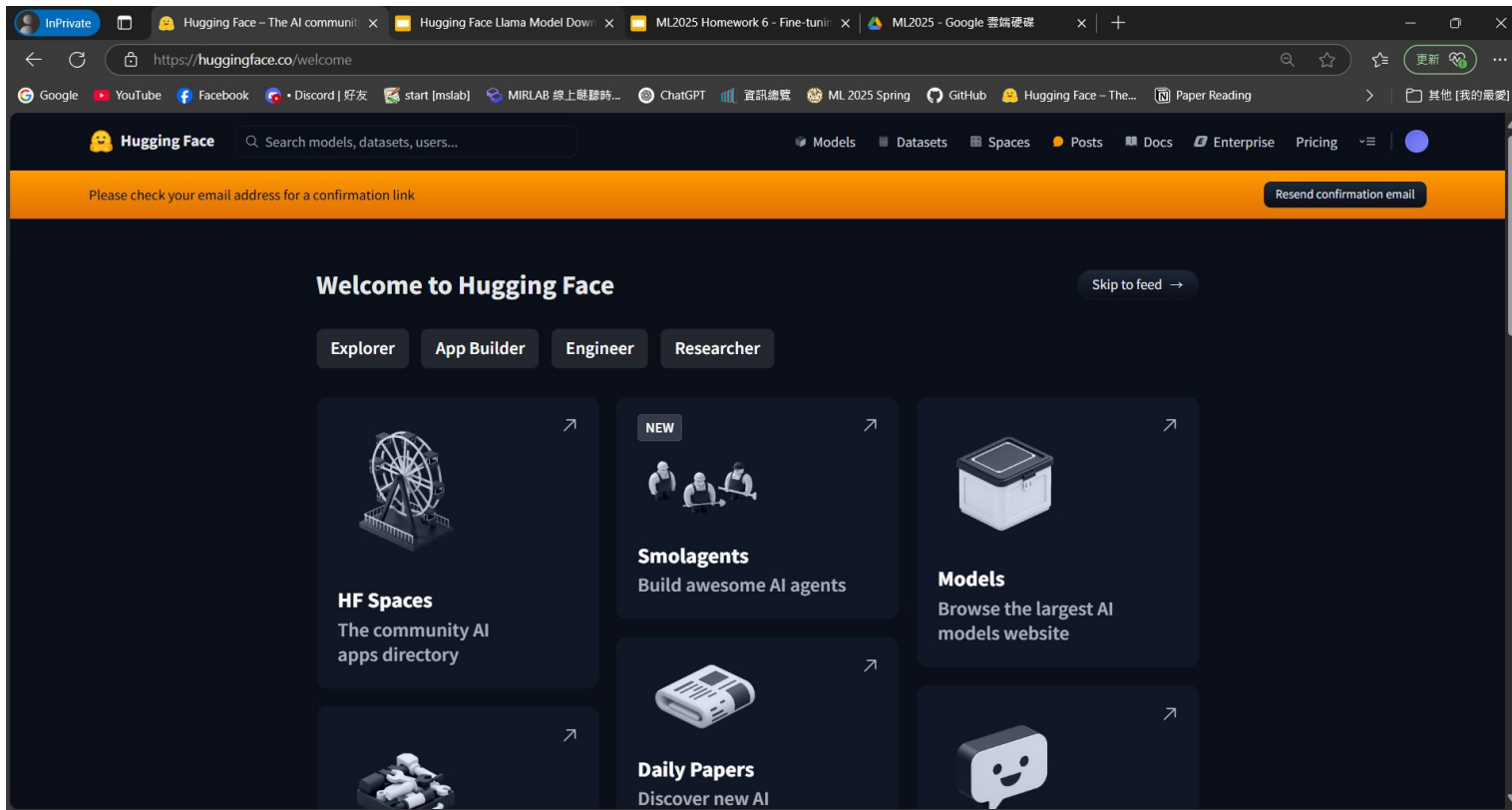
Homepage (optional): Homepage

AI & ML interests (optional): AI & ML interests

☒ I have read and agree with the [Terms of Service](#) and the [Code of Conduct](#)

Create Account

Create Huggingface Account



Submit Llama Access Request

Submit Llama Access Request

The screenshot shows a web browser window with the Hugging Face website. The address bar displays `https://huggingface.co/email_confirmation`. The page has a dark theme. On the left, a sidebar menu is open, showing a search bar with the text "llama3.2-1b-instruct". Below the search bar, the "Models" section is expanded, listing several models, with "meta-llama/Llama-3.2-1B-Instruct" highlighted. Other sections like "Datasets" and "Spaces" are also visible. The main content area features a large yellow emoji of a smiling face with its hands clasped. Below the emoji, the text reads: "Confirmation email has been sent, please check your mailbox". A button labeled "Resend confirmation email" is located in the top right corner of the main content area. The footer of the page contains links for "Company", "Resources", and "Social".

Hugging Face | Search: llama3.2-1b-instruct

Models

- meta-llama/Llama-3.2-1B-Instruct
- bartowski/Llama-3.2-1B-Instruct-GGUF
- unsloth/Llama-3.2-1B-Instruct
- unsloth/Llama-3.2-1B-Instruct-GGUF
- Mungert/Llama-3.2-1B-Instruct-GGUF
- meta-llama/Llama-3.2-1B-Instruct-SpinQuant_INT4_E08
- See 1508 model results for "llama3.2-1b-instruct"

Datasets

- meta-llama/Llama-3.2-1B-Instruct-ovals
- HuggingFaceH4/Llama-3.2-1B-Instruct-best-of-N-completions
- HuggingFaceH4/Llama-3.2-1B-Instruct-beam-search-completions
- See 94 dataset results for "llama3.2-1b-instruct"

Spaces

- Nymbo/Llama-3.2-1B-Instruct
- GokuRajaR/Llama_3.2_1B_Instruct_LitServe
- hamsomp3/tecnicas2024-llama3.2-1b-instruct
- See 7 Space results for "llama3.2-1b-instruct"

Use Full-text search →

Confirmation email has been sent, please check your mailbox

[Resend confirmation email](#)

Company

- About
- Brand assets
- Terms of service

Resources

- Learn
- Documentation
- Blog

Social

- GitHub
- Twitter
- LinkedIn

Submit Llama Access Request

The screenshot shows the Hugging Face interface for the `meta-llama/Llama-3.2-1B-Instruct` model. The page is in dark mode. At the top, the browser tabs and address bar are visible, showing the URL `https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct`. The Hugging Face logo and search bar are at the top left. The model name `meta-llama/Llama-3.2-1B-Instruct` is prominently displayed, along with its statistics (likes, follows) and various tags like `Text Generation`, `Transformers`, `Safetensors`, `PyTorch`, `8 languages`, `llama`, `facebook`, `meta`, `llama-3`, `conversational`, `text-generation-inference`, `arxiv:2204.05149`, and `arxiv:2405.16406`. The license is listed as `llama3.2`. Below the model name, there are tabs for `Model card`, `Files and versions`, and `Community`. The `Model card` tab is selected, showing a section titled `You need to agree to share your contact information to access this model`. This section contains the `LLAMA 3.2 COMMUNITY LICENSE AGREEMENT` text, which includes details about the license and a link to the full agreement. A red box highlights the `Expand to review and access` button at the bottom of this section. On the right side of the page, there is a section for `Downloads last month` showing `2,940,986` downloads, a line graph, and sections for `Safetensors` and `Inference Providers`.

meta-llama/Llama-3.2-1B-Instruct like 845 Follow Meta Llama 34k

Text Generation Transformers Safetensors PyTorch 8 languages llama facebook meta llama-3 conversational text-generation-inference arxiv:2204.05149 arxiv:2405.16406

License: llama3.2

Model card Files and versions Community 82

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

Downloads last month
2,940,986

Safetensors Model size 1.24B params Tensor type BF16

Inference Providers SambaNova +1

Text Generation Examples

Input a message to start chatting with meta-llama/Llama-3.2-1B-Instruct.

Submit Llama Access Request

The screenshot shows a web browser window with the URL <https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct>. The browser's address bar and tabs are visible at the top. The main content area displays a form for submitting an access request. The form includes fields for First Name (Bo-Han), Last Name (Feng), Date of birth (2003/07/19), Country (Taiwan), Affiliation (National Taiwan University), and Job title (Student). A checkbox is checked, indicating acceptance of the terms of the license. A red rectangle highlights the 'Submit' button at the bottom of the form.

Reporting violations of the Acceptable Use Policy or unlicensed uses of Llama 3.2: LlamaUseReport@meta.com

By agreeing you accept to share your contact information (email and username) with the repository authors.

First Name
Bo-Han

Last Name
Feng

Date of birth
2003/07/19

Country
Taiwan

Affiliation
National Taiwan University

Job title
Student

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

Submit Llama Access Request

The screenshot shows the Hugging Face interface for the `meta-llama/Llama-3.2-1B-Instruct` model. The browser address bar shows the URL `https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct`. The page header includes the Hugging Face logo and navigation links for Models, Datasets, Spaces, Posts, Docs, Enterprise, and Pricing. The model card displays the name `meta-llama/Llama-3.2-1B-Instruct` with 845 likes and 34k followers. It lists various tags such as Text Generation, Transformers, Safetensors, PyTorch, 8 languages, llama, facebook, meta, llama-3, conversational, text-generation-inference, and two arXiv IDs. The license is `llama3.2`. The Model card tab is selected, showing a section titled "You need to agree to share your contact information to access this model" with a link to the Meta Privacy Policy. Below this is the "LLAMA 3.2 COMMUNITY LICENSE AGREEMENT" section, which includes the release date (September 25, 2024) and definitions for "Agreement", "Documentation", and "Licensee". A blue box states: "Your request to access this repository has been submitted and is awaiting a review from the repository authors. You can check the status of all your access requests in your settings." Another blue box states: "If approved, you'll get access to all 14 repositories in the Meta's Llama 3.2 language models & evals Gating Group Collection, including this one." On the right side, there is a section for Downloads last month (2,940,986) with a line graph, a Safetensors section showing model size (1.24B params) and tensor type (BF16), and an Inference Providers section listing SambaNova and others. At the bottom right, there is a chat interface with a text input field and a button to start chatting.

meta-llama/Llama-3.2-1B-Instruct

Text Generation Transformers Safetensors PyTorch 8 languages llama facebook meta llama-3 conversational text-generation-inference arxiv:2204.05149 arxiv:2405.16406

License: llama3.2

Model card Files and versions Community 82

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...)

Your request to access this repository has been submitted and is awaiting a review from the repository authors. You can check the status of all your access requests in your [settings](#).

If approved, you'll get access to all 14 repositories in the [Meta's Llama 3.2 language models & evals Gating Group Collection](#), including this one.

Downloads last month
2,940,986

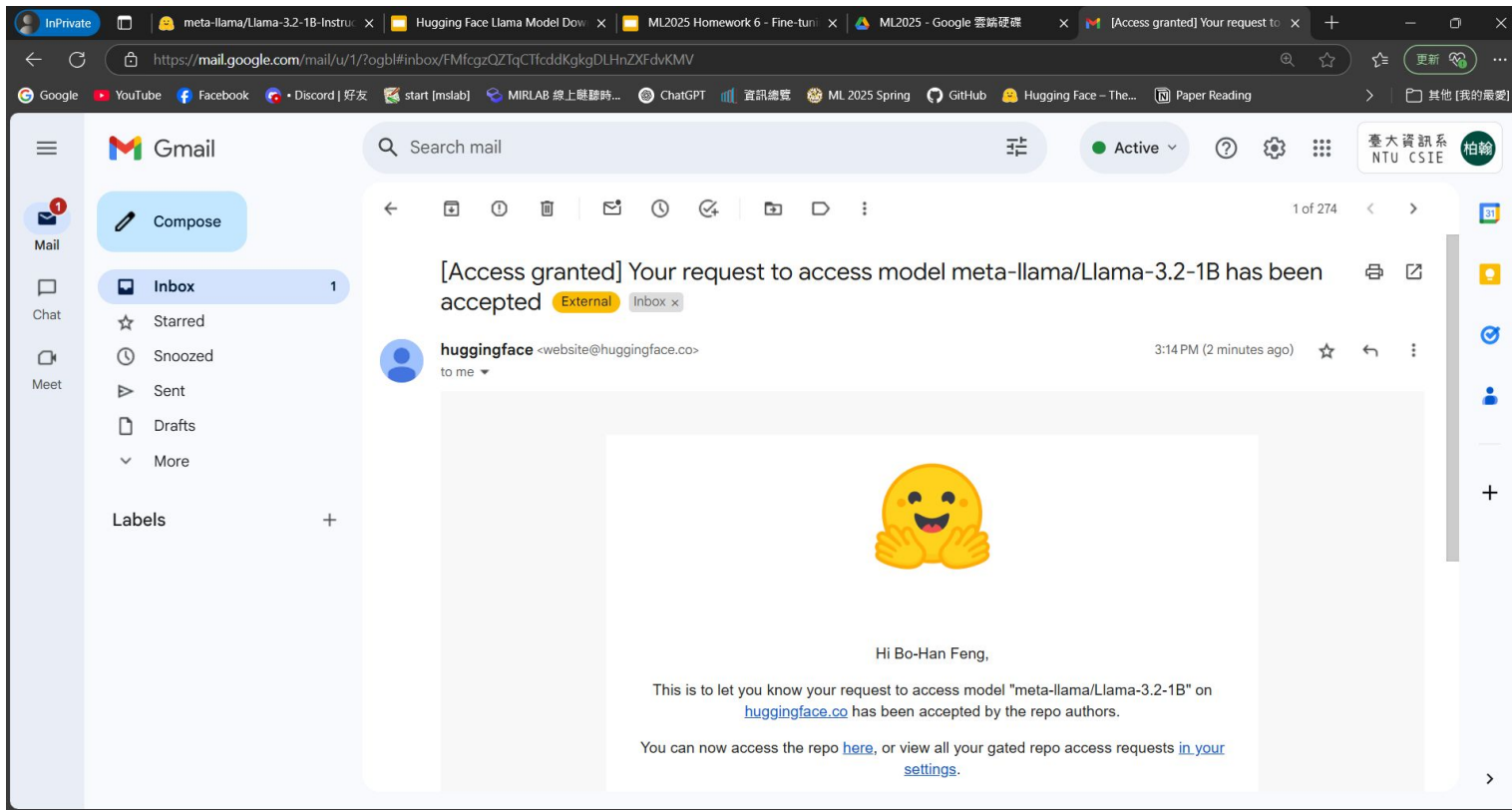
Safetensors Model size 1.24B params Tensor type BF16

Inference Providers SambaNova +1

Text Generation Examples

Input a message to start chatting with meta-llama/Llama-3.2-1B-Instruct.

Submit Llama Access Request



Create Access Token

Create Access Token

The screenshot shows the Hugging Face interface for the `meta-llama/Llama-3.2-1B-Instruct` model. The page includes a model card with information about the Llama 3.2 collection, a sidebar with navigation options, and a user profile menu on the right. The `Access Tokens` option in the profile menu is highlighted with a red box.

Model Information

The Llama 3.2 collection of multilingual large language models (LLMs) is a collection of pretrained and instruction-tuned generative models in 1B and 3B sizes (text in/text out). The Llama 3.2 instruction-tuned text only models are optimized for multilingual dialogue use cases, including agentic retrieval and summarization tasks. They outperform many of the available open source and closed chat models on common industry benchmarks.

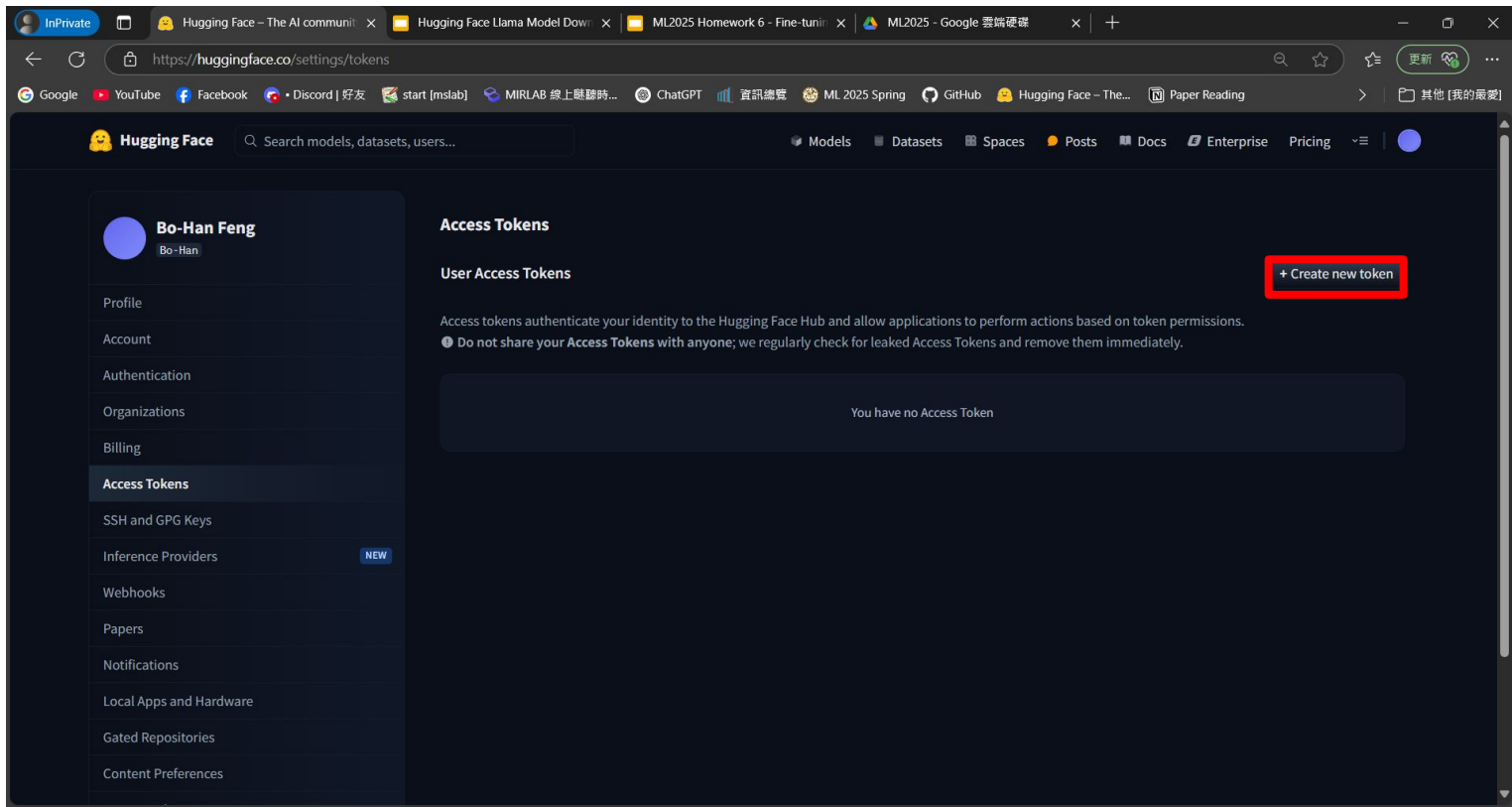
Model Developer: Meta

Model Architecture: Llama 3.2 is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety.

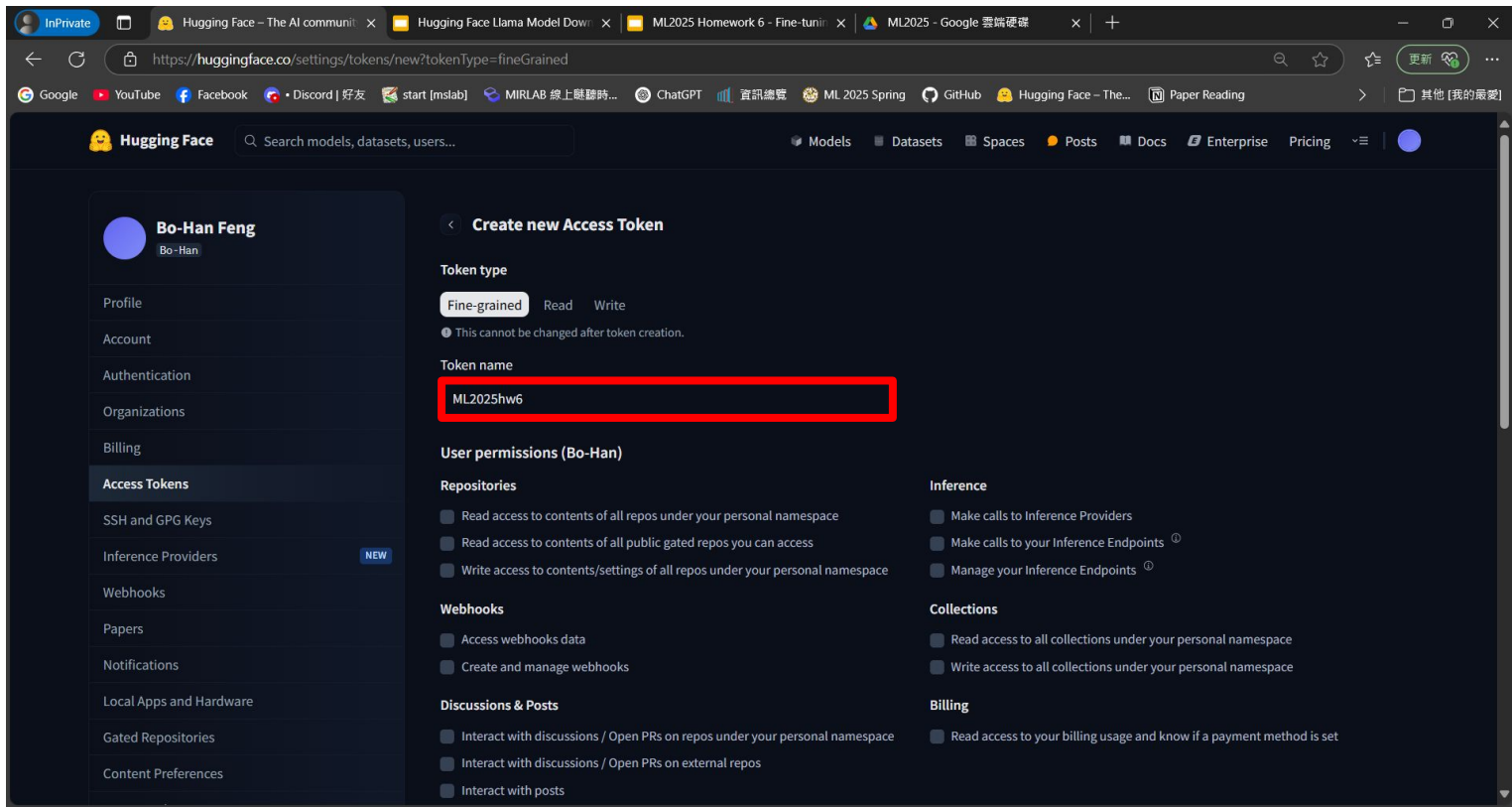
User Profile Menu:

- Profile: Bo-Han
- Notifications: Inbox (0)
- + New Model
- + New Dataset
- + New Space
- + New Collection
- Create organization
- Usage Quota
- Private Storage: 0 GB/100 GB
- Zero GPU: 0/5 min
- Inference Usage: \$0.00 / \$0.10
- Subscribe to PRO
- Settings
- Access Tokens**
- Billing
- Sign Out

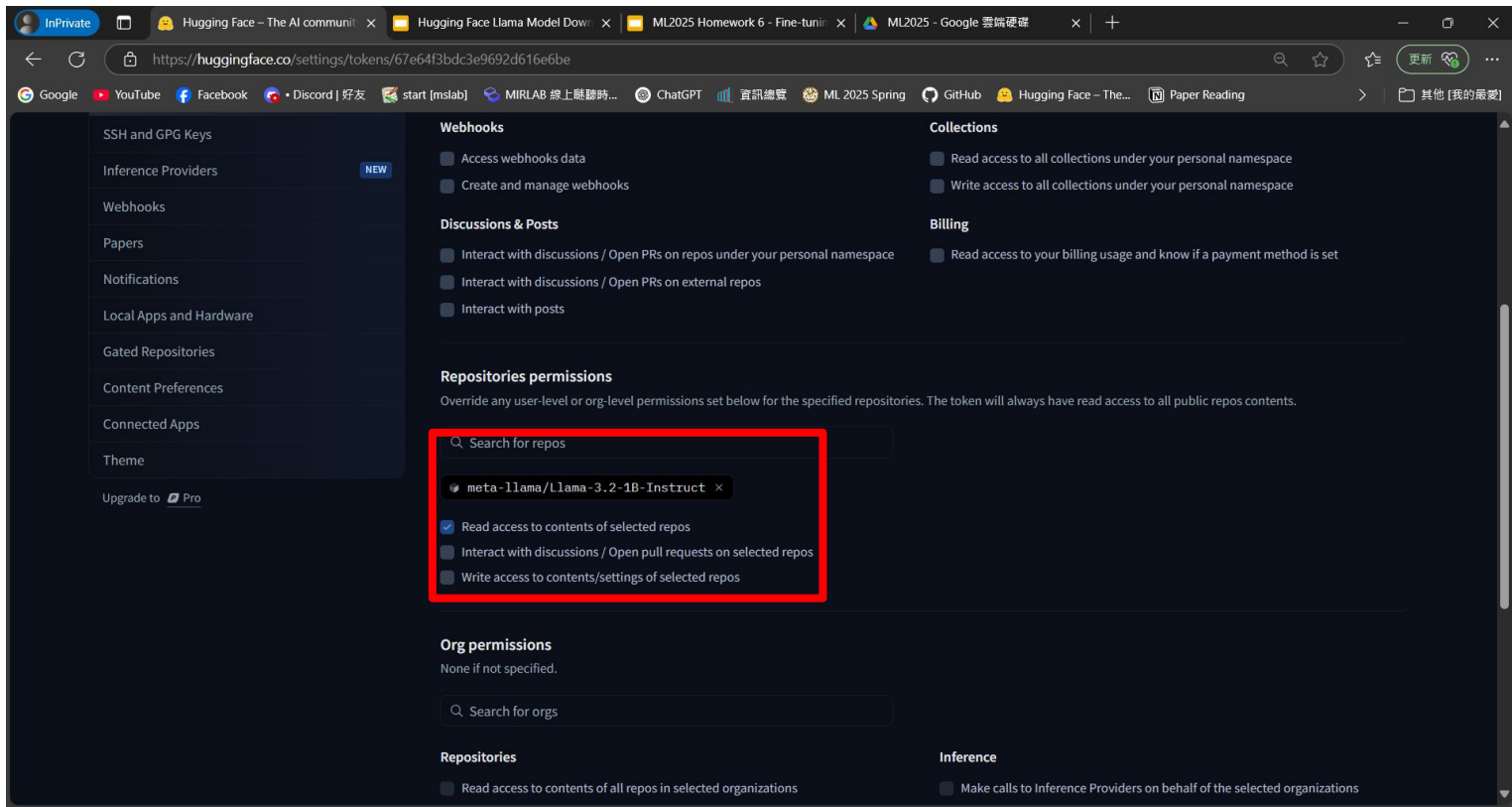
Create Access Token



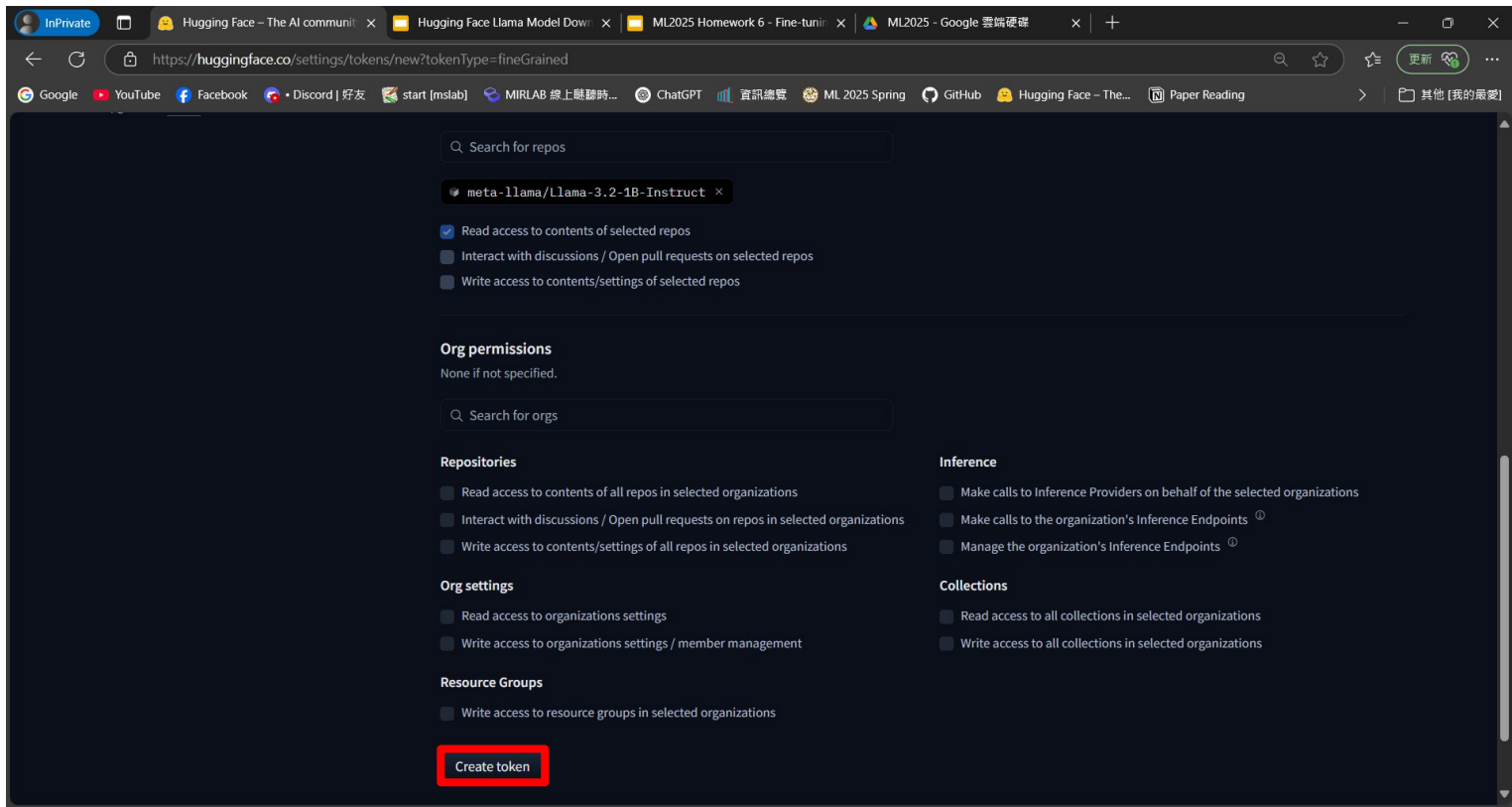
Create Access Token



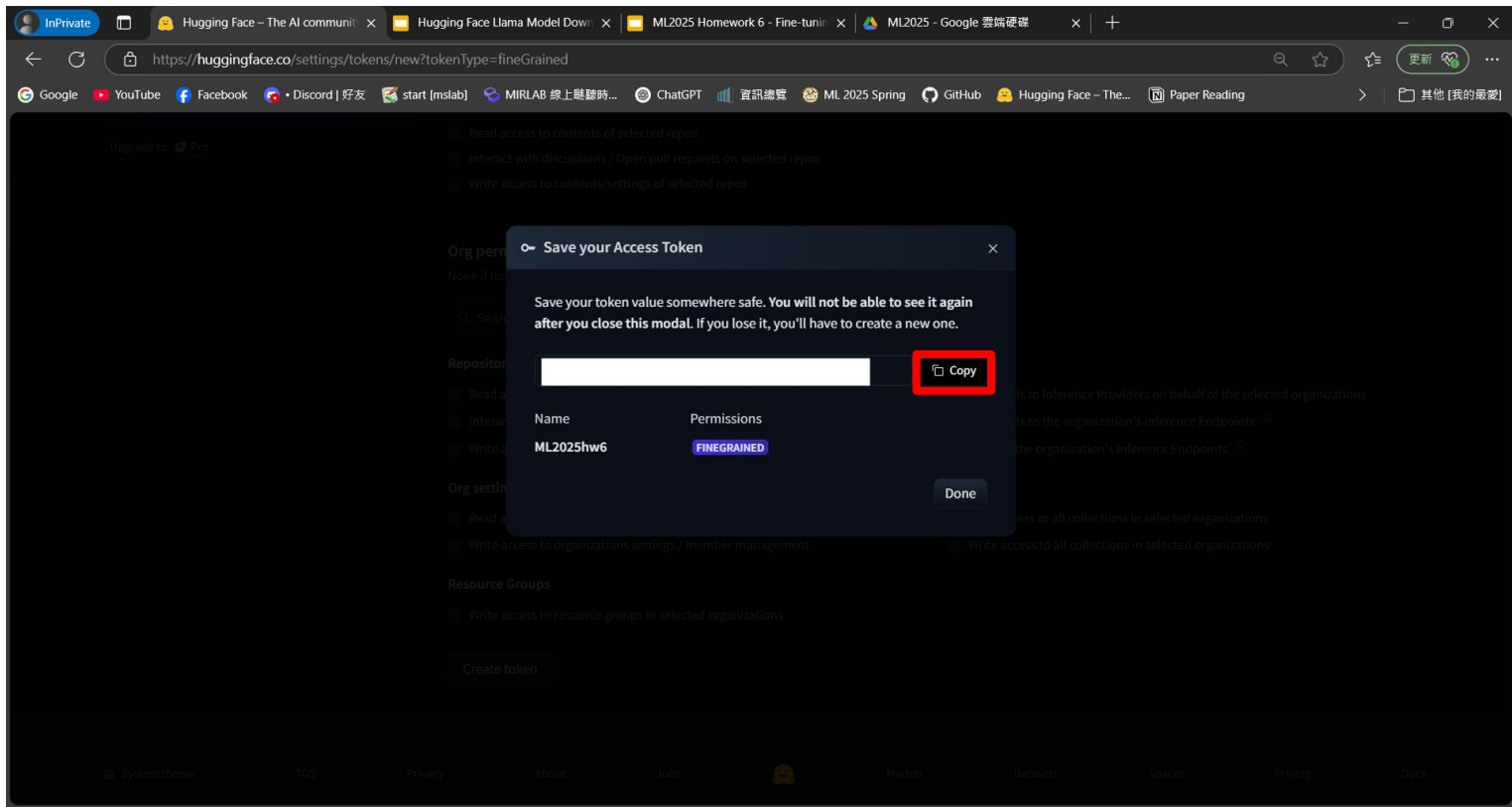
Create Access Token



Create Access Token



Create Access Token



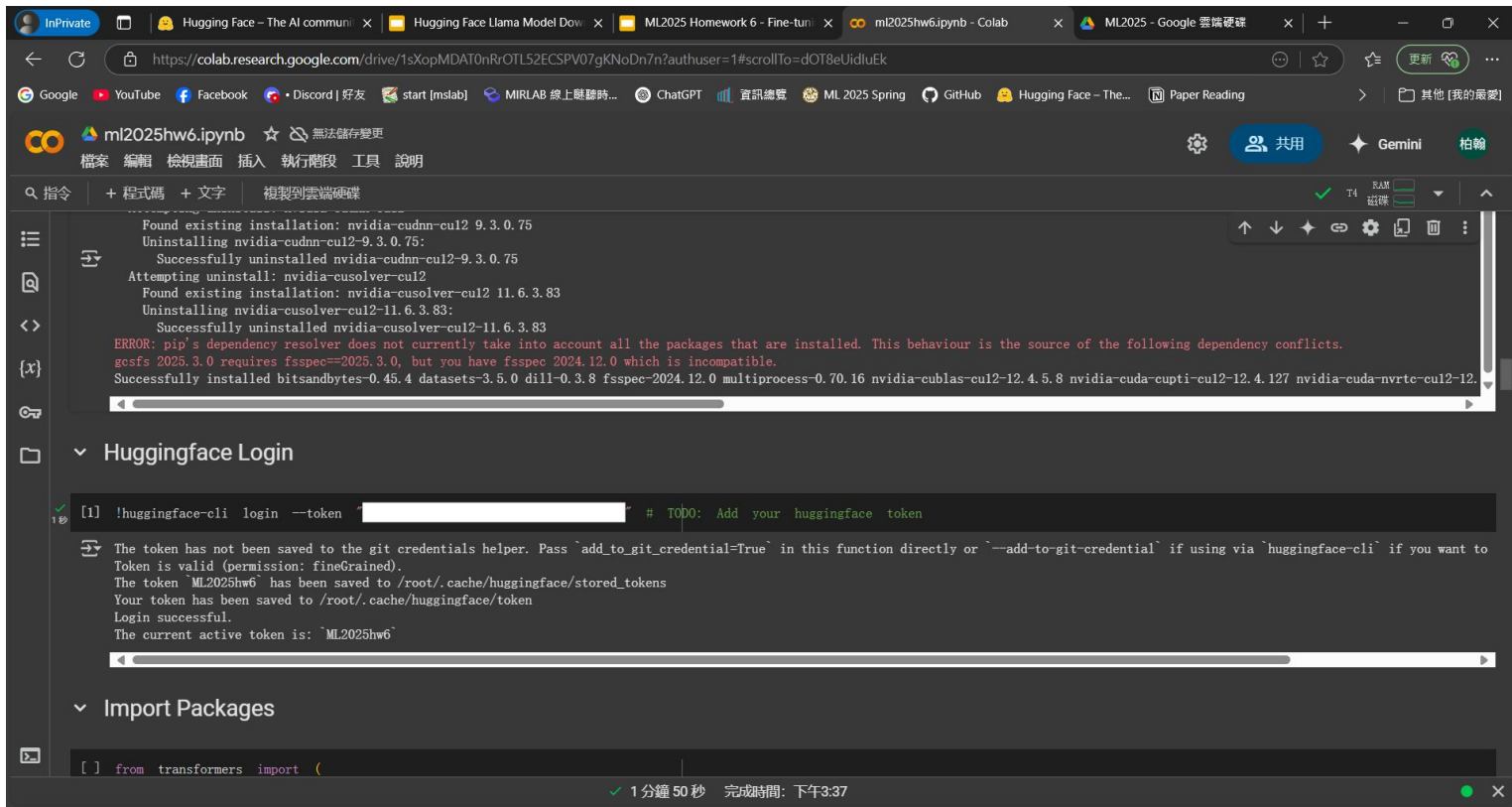
Create Access Token

The screenshot displays the Hugging Face 'Access Tokens' management interface. On the left, a sidebar lists various settings: Profile, Account, Authentication, Organizations, Billing, Access Tokens (highlighted), SSH and GPG Keys, Inference Providers (marked as NEW), Webhooks, Papers, Notifications, Local Apps and Hardware, Gated Repositories, and Content Preferences. The main content area is titled 'Access Tokens' and includes a '+ Create new token' button. Below this, a section explains that access tokens authenticate identity and allow applications to perform actions based on permissions, with a warning not to share tokens. A table lists the existing tokens:

Name	Value	Last Refreshed Date	Last Used Date	Permissions
ML2025hw6	hf_...AoOH	1 minute ago	-	FINEGRAINED

Login Hugging Face and Download Llama in Colab

Login Hugging Face and Download Llama in Colab



The screenshot shows a Google Colab notebook with the following content:

Terminal Output:

```
Found existing installation: nvidia-cudnn-cu12 9.3.0.75
Uninstalling nvidia-cudnn-cu12-9.3.0.75:
Successfully uninstalled nvidia-cudnn-cu12-9.3.0.75
Attempting uninstall: nvidia-cusolver-cu12
Found existing installation: nvidia-cusolver-cu12 11.6.3.83
Uninstalling nvidia-cusolver-cu12-11.6.3.83:
Successfully uninstalled nvidia-cusolver-cu12-11.6.3.83
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency conflicts.
gcsfs 2025.3.0 requires fsspec==2025.3.0, but you have fsspec 2024.12.0 which is incompatible.
Successfully installed bitsandbytes-0.45.4 datasets-3.5.0 dill-0.3.8 fsspec-2024.12.0 multiprocessing-0.70.16 nvidia-cublas-cu12-12.4.5.8 nvidia-cuda-cupti-cu12-12.4.127 nvidia-cuda-nvrtc-cu12-12.4.127 nvidia-cuda-runtime-cu12-12.4.127 nvidia-cudnn-cu12-9.3.0.75 nvidia-cusolver-cu12-11.6.3.83 nvidia-cuxx-cu12-12.4.127
```

Huggingface Login

```
[1] !huggingface-cli login --token [REDACTED] # TODO: Add your huggingface token
```

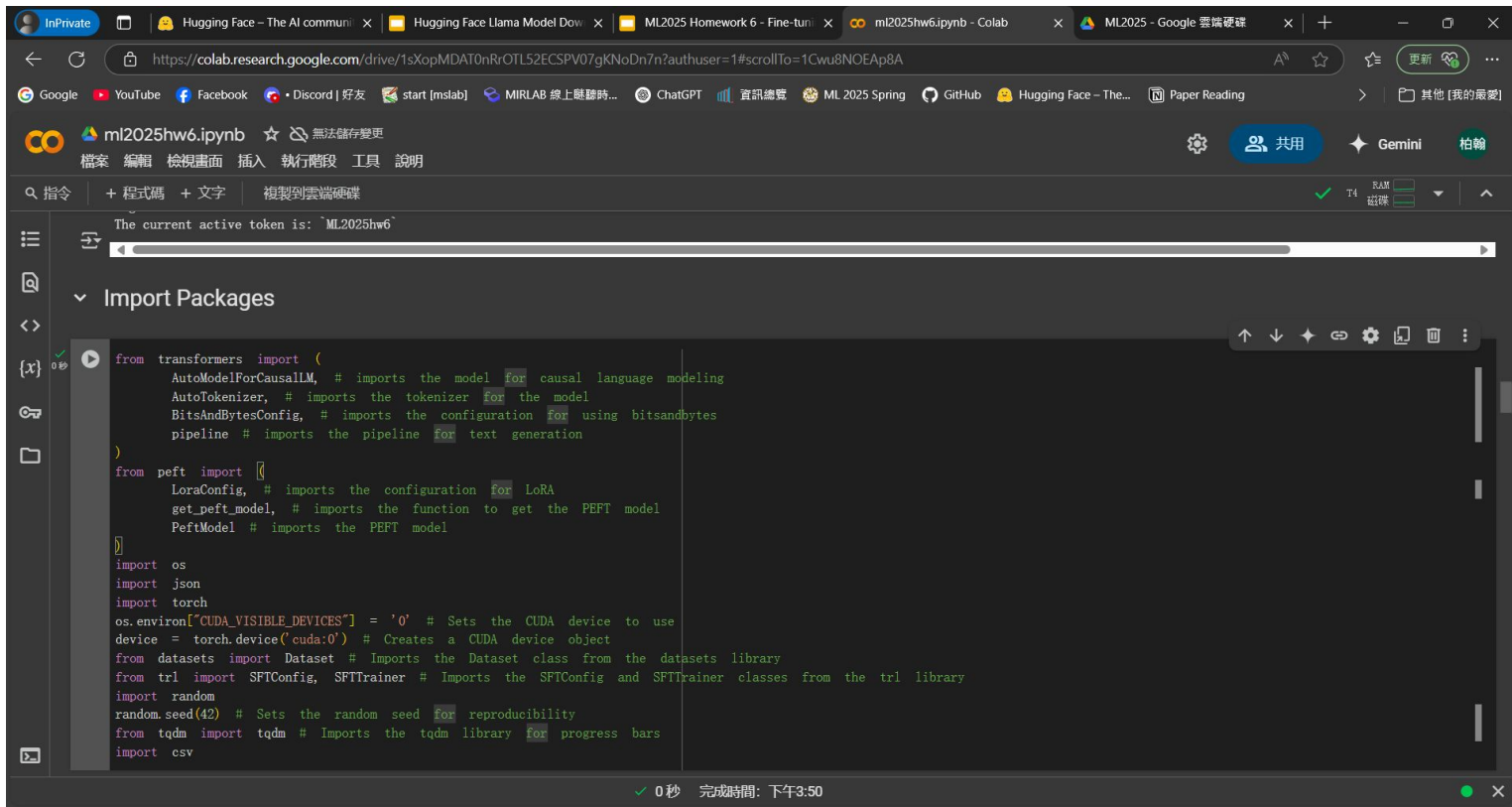
The token has not been saved to the git credentials helper. Pass `add_to_git_credential=True` in this function directly or `--add-to-git-credential` if using via `huggingface-cli` if you want to. Token is valid (permission: fineGrained). The token `ML2025hw6` has been saved to `/root/.cache/huggingface/stored_tokens`. Your token has been saved to `/root/.cache/huggingface/token`. Login successful. The current active token is: `ML2025hw6`.

Import Packages

```
[ ] from transformers import (
```

Footer: 1 分鐘 50 秒 完成時間: 下午3:37

Login Hugging Face and Download Llama in Colab



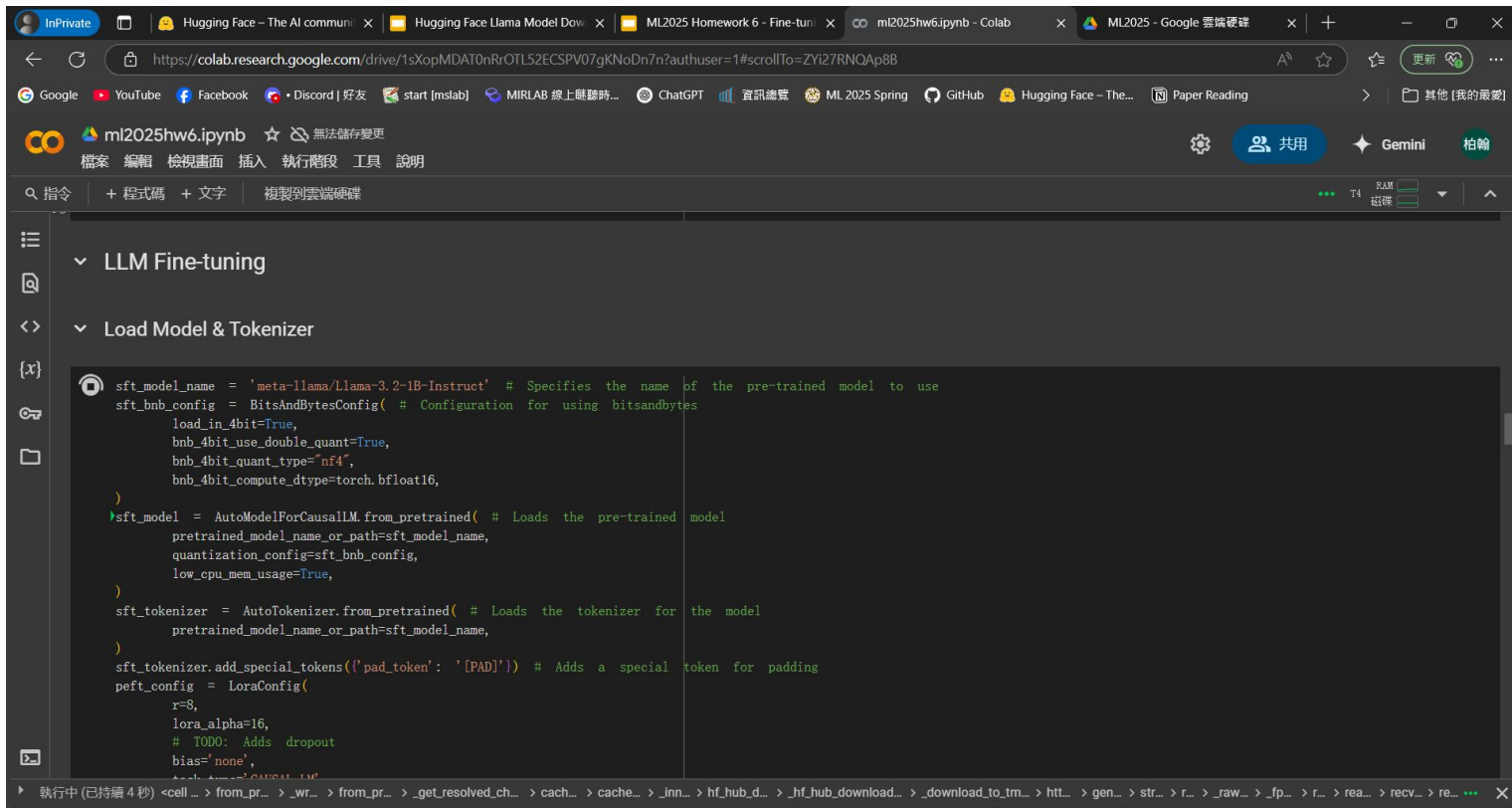
The current active token is: `ML2025hw6`

Import Packages

```
from transformers import (
    AutoModelForCausalLM, # imports the model for causal language modeling
    AutoTokenizer, # imports the tokenizer for the model
    BitsAndBytesConfig, # imports the configuration for using bitsandbytes
    pipeline # imports the pipeline for text generation
)
from peft import (
    LoraConfig, # imports the configuration for LoRA
    get_peft_model, # imports the function to get the PEFT model
    PeftModel # imports the PEFT model
)
import os
import json
import torch
os.environ["CUDA_VISIBLE_DEVICES"] = '0' # Sets the CUDA device to use
device = torch.device('cuda:0') # Creates a CUDA device object
from datasets import Dataset # Imports the Dataset class from the datasets library
from trl import SFTConfig, SFTTrainer # Imports the SFTConfig and SFTTrainer classes from the trl library
import random
random.seed(42) # Sets the random seed for reproducibility
from tqdm import tqdm # Imports the tqdm library for progress bars
import csv
```

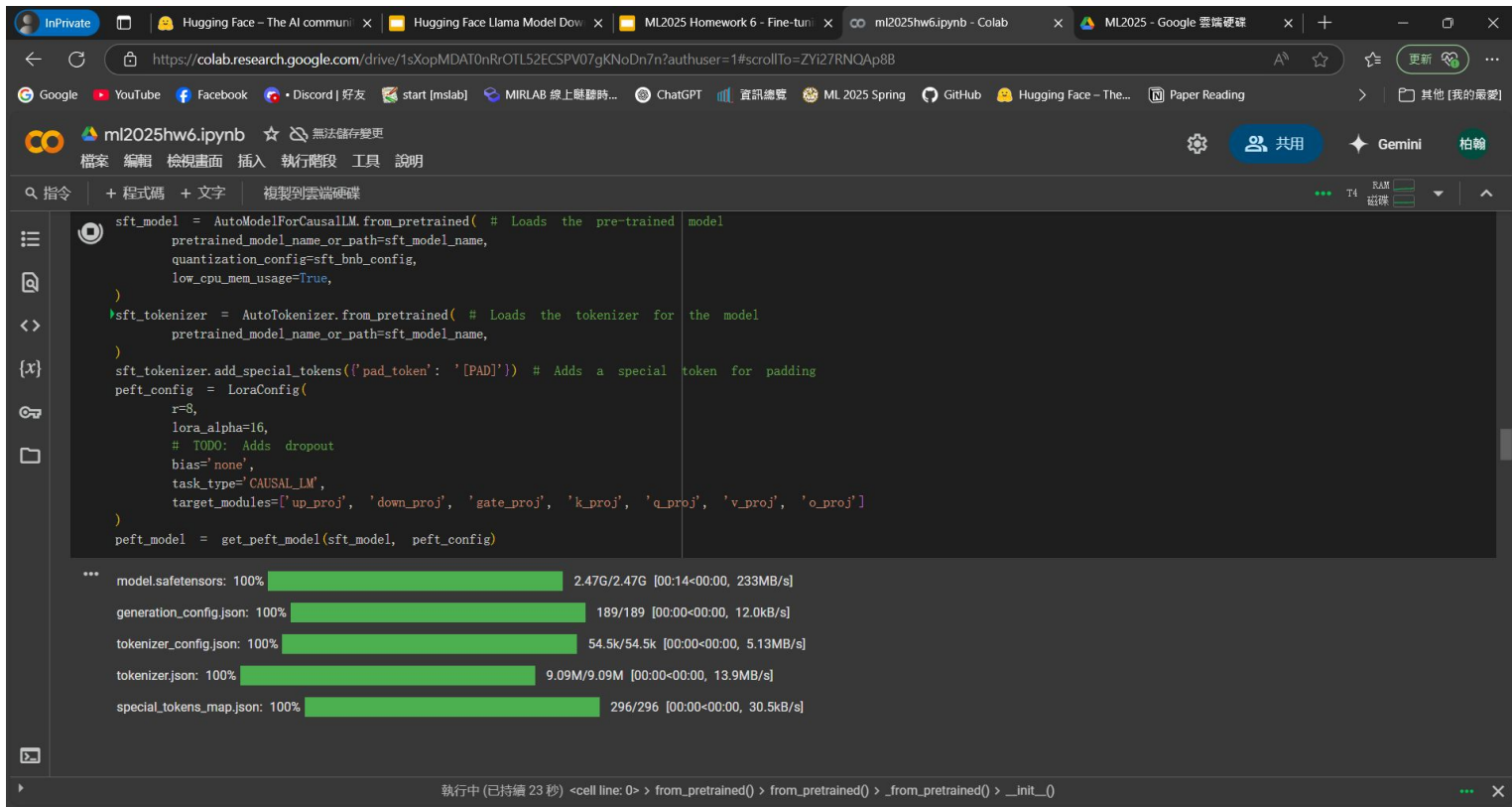
0 秒 完成時間: 下午3:50

Login Hugging Face and Download Llama in Colab



```
sft_model_name = 'meta-llama/Llama-3.2-1B-Instruct' # Specifies the name of the pre-trained model to use
sft_bnb_config = BitsAndBytesConfig( # Configuration for using bitsandbytes
    load_in_4bit=True,
    bnb_4bit_use_double_quant=True,
    bnb_4bit_quant_type='nf4',
    bnb_4bit_compute_dtype=torch.bfloat16,
)
sft_model = AutoModelForCausalLM.from_pretrained( # Loads the pre-trained model
    pretrained_model_name_or_path=sft_model_name,
    quantization_config=sft_bnb_config,
    low_cpu_mem_usage=True,
)
sft_tokenizer = AutoTokenizer.from_pretrained( # Loads the tokenizer for the model
    pretrained_model_name_or_path=sft_model_name,
)
sft_tokenizer.add_special_tokens({'pad_token': '[PAD]'}) # Adds a special token for padding
peft_config = LoraConfig(
    r=8,
    lora_alpha=16,
    # TODO: Adds dropout
    bias='none',
    torch_dtype=torch.bfloat16)
```

Login Hugging Face and Download Llama in Colab



The screenshot shows a Google Colab notebook interface. The browser tabs at the top include 'InPrivate', 'Hugging Face - The AI community', 'Hugging Face Llama Model Download', 'ML2025 Homework 6 - Fine-tuning', 'ml2025hw6.ipynb - Colab', 'ML2025 - Google 雲端硬碟', and others. The notebook's URL is <https://colab.research.google.com/drive/1sXopMDAT0nRrOTL52ECSPV07gKNoDn7n?authuser=1#scrollTo=ZYI27RNQAp8B>. The notebook title is 'ml2025hw6.ipynb'. The code in the cell is as follows:

```
sft_model = AutoModelForCausalLM.from_pretrained( # Loads the pre-trained model
    pretrained_model_name_or_path=sft_model_name,
    quantization_config=sft_bnb_config,
    low_cpu_mem_usage=True,
)
sft_tokenizer = AutoTokenizer.from_pretrained( # Loads the tokenizer for the model
    pretrained_model_name_or_path=sft_model_name,
)
sft_tokenizer.add_special_tokens({'pad_token': '[PAD]'}) # Adds a special token for padding
peft_config = LoraConfig(
    r=8,
    lora_alpha=16,
    # TODO: Adds dropout
    bias='none',
    task_type='CAUSAL_LM',
    target_modules=['up_proj', 'down_proj', 'gate_proj', 'k_proj', 'q_proj', 'v_proj', 'o_proj']
)
peft_model = get_peft_model(sft_model, peft_config)
```

Below the code, the execution progress is shown with green progress bars and text:

```
***
model.safetensors: 100% 2.47G/2.47G [00:14<00:00, 233MB/s]
generation_config.json: 100% 189/189 [00:00<00:00, 12.0kB/s]
tokenizer_config.json: 100% 54.5k/54.5k [00:00<00:00, 5.13MB/s]
tokenizer.json: 100% 9.09M/9.09M [00:00<00:00, 13.9MB/s]
special_tokens_map.json: 100% 296/296 [00:00<00:00, 30.5kB/s]
```

The status bar at the bottom indicates the execution is ongoing: '執行中 (已持續 23 秒) <cell line: 0> > from_pretrained() > from_pretrained() > _from_pretrained() > __init__()'. On the right side of the status bar, there are indicators for RAM and disk usage, showing 'T4' and '磁碟' respectively.