# Edge AI ✨

> 🤖 AI processing on local devices

OAD brian_li

# Agenda

- Edge AI
- Open Source Model
- Ollama
- Open WebUI
- Unsloth
- LLM Tools
- Taiwan Llama
- Live Demo

# Edge AI

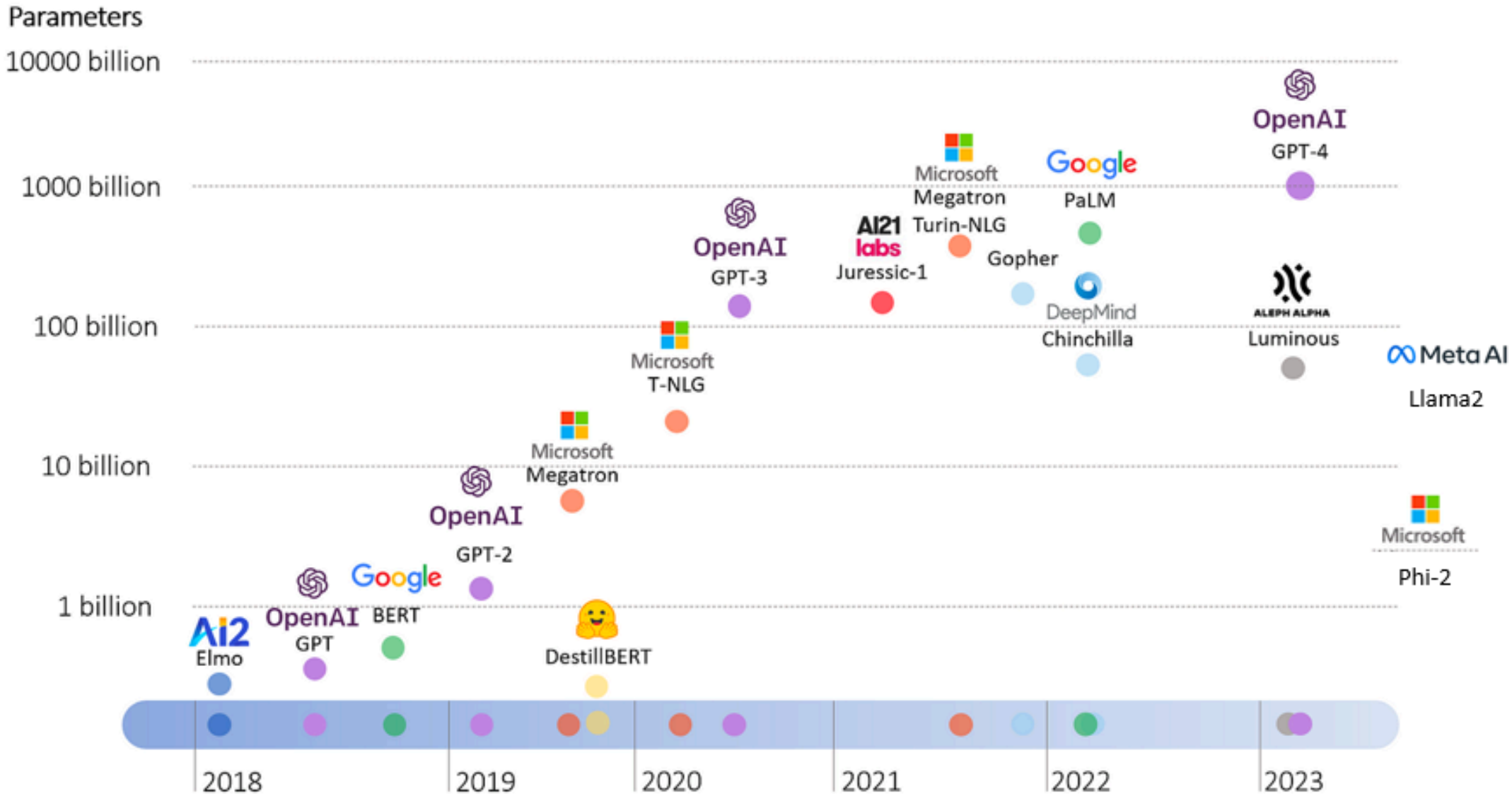
**允許數據在本地進行處理，而不需傳輸到雲端服務**

- 低延遲：即時反應
- 安全性：本地處理數據
- 可靠性：網絡中斷仍可運行

⚠️ 性能表現依賴本地設備

# Open Source vs Close Models

| 特點 | 開源模型 | 閉源模型 |
| --- | --- | --- |
| 可用性 | 免費、公開 | 需購買或授權 |
| 擴充性 | 可自由修改、擴展 | 依賴提供者 |
| 透明度 | 高，公開源碼 | 低，不公開 |
| 安全性 | 可檢查漏洞 | 依賴提供者 |
| 社群支援 | 大量社群、更新快 | 無 |
| 依賴性 | 不依賴單一提供者 | 高度依賴提供者 |
| 舉例 | `llama3` `gemma2` `phi3` | `gpt4o` `sonnet` |


Parameters
10000 billion
1000 billion
100 billion
10 billion
1 billion
Elmo
OpenAI
GPT
Google
BERT
OpenAI
GPT-2
Microsoft
Megatron
DestillBERT
Microsoft
T-NLG
OpenAI
GPT-3
AI21 labs
Jurassic-1
Microsoft
Megatron
Turin-NLG
Gopher
Google
PaLM
DeepMind
Chinchilla
OpenAI
GPT-4
ALEPH ALPHA
Luminous
Meta AI
Llama2
Microsoft
Phi-2
2018
2019
2020
2021
2022
2023

# Ollama

https://ollama.com/

- 一個開源工具
- 允許在本地運行各種開源 LLM
- 支援 GPU 加速 (如果有)
- 允許微調模型或創建自定義模型
- 支援 CLI 操作
- 提供 REST API

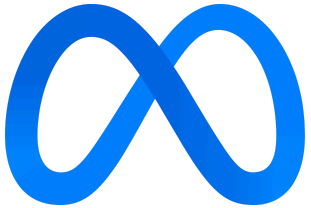
# Risks of Close Model

OpenAI `GPT4o` 、Anthropic `Sonnet`

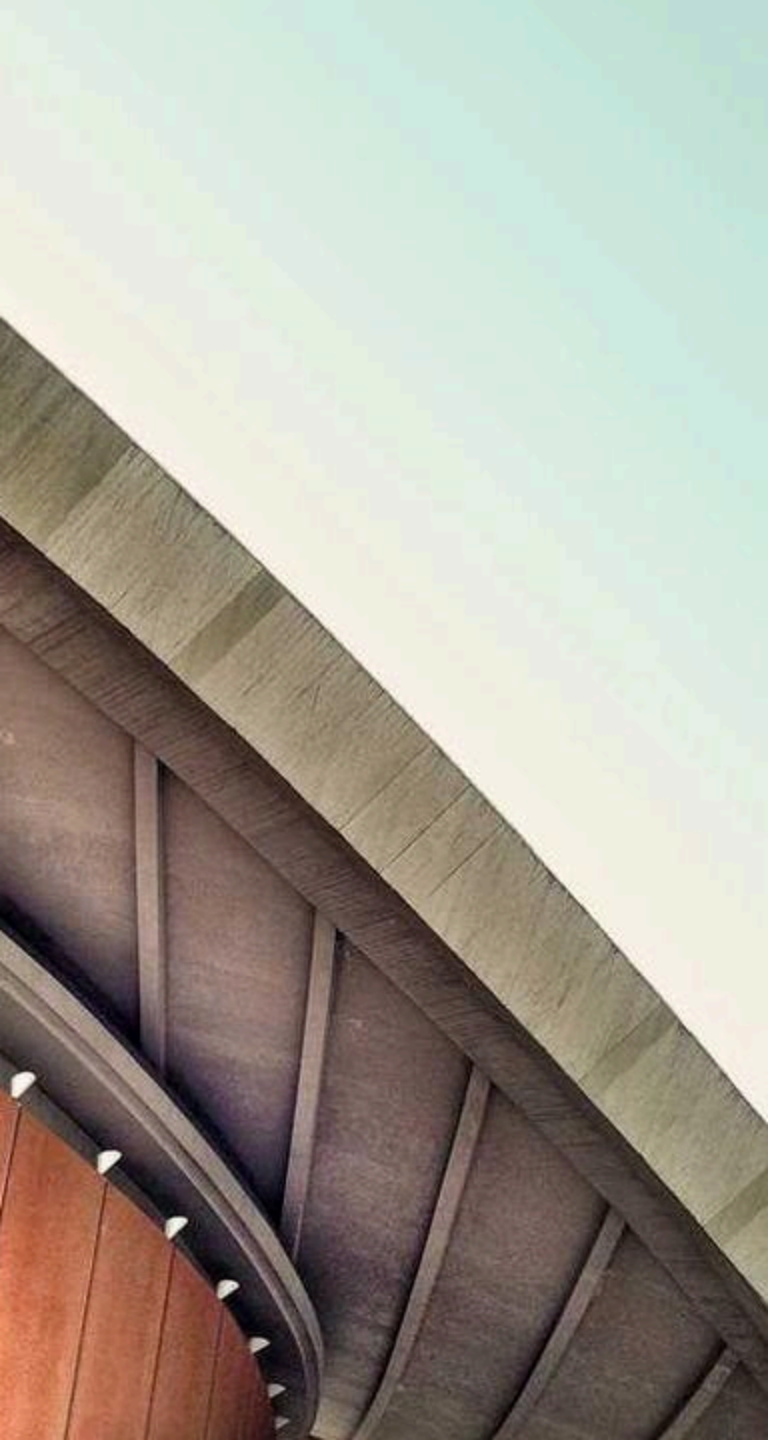
- **無法避免資料訓練**
  Prompt
  Document
- **完全依賴網路**
  中斷
  延遲

> ⚠️服務商持有控制權

# 常見開源模型

|  |  |  |  |  |
|---|---|---|---|---|
| Meta | Mistrel AI | Google | Microsoft | Alibaba |
| **llama3** | **mistral** | **gemma2** | **phi3** | **qwen2** |
| `8B` `70B` | `7B` | `9B` `27B` | `3B` `14B` | `0.5B` `1.5B` `7B` `72B` |

> `B` = `Billion` 十億，代表模型參數(Parameters)數量，例如 GPT-3 為 `175B`



# Ollama - CLI

```
# 檢視版本
ollama -v

# 下載模型，如需指定參數請用例如 llama3:8b
ollama pull llama3

# 與 AI 互動
ollama run llama3 "Why is the sky blue?"

# 其他指令查詢
ollama -h
```

> 其他指令操作可參考線上說明

# Ollama - API

```
# 使用 curl 呼叫 API

curl -X POST http://localhost:11434/api/generate -d '{
  "model": "llama3",
  "prompt":"Why is the sky blue?"
 }'
```

- 詳細 API 端點資訊請 參考文件
- 也可使用 `Postman` 或 `*.http` 測試

> 💡應用程式可透過 API 與 model 互動

# Open WebUI

https://openwebui.com/

- 支援 Ollama = 可用開源 LLM
- 支援 OpenAI = 可用 GPT4o
- 類似 ChatGPT，學習曲線低
- 支援檔案上傳、檔案管理保存
- 支援RAG，可搜尋網頁或檔案
- 可自訂 AI 模型(類似GPTs)
- 支援語音輸入與輸出
- 支援多模型提問
- 可上傳自訓練模型(GGUF)

New Chat

Workspace

Search

OpenAI / GPT 4

Set as default

SUCCESS Open WebUI - On a mission to build the best open-source AI user interface.

# OpenAI / GPT 4

## How can I help you today?

Suggested

**Help me study**
vocabulary for a college entrance exam
Prompt

**Give me ideas**
for what to do with my kids' art
Prompt

**Overcome procrastination**
give me tips
Prompt

**Tell me a fun fact**
about the Roman Empi
Prompt

Send a Message

Timothy J. Baek

LLMs can make mistakes. Verify important information.

12

# RAG vs Fine-Tuning

👨‍🎨 小明 收藏了整套哈利波特，完整閱讀過，大概知道章節與內容

👩 小美 熟讀了整套哈利波特，內容到背如流，後來將書捐贈出售

👱 小華 請教 👨‍🎨 小明 哈利波特內容， 👨‍🎨 小明 快速查閱找到答案

👱 小華 請教 👩 小美 哈利波特內容， 👩 小美 不加思索立刻回答

👨‍🎨 小明 就是 **RAG** (Retrieval-Augmented Generation)

👩 小美 就是 **Fine-Tuning**

> 後來 J.K.羅琳 決定推出哈利波特新系列.... 👨‍🎨 vs 👩將會?

# unsloth

https://unsloth.ai/

- 快速微調和訓練 LLM 的開源工具
- 訓練速度較其它工具快，使用記憶體低
- 支援 `NVIDIA` `AMD` `Intel` 等 GPU
- 在 Google Colab 即可免費訓練*
- GGUF 可導入 OpenWebUI 使用

> *⚠️免費僅限單CPU，模型不可過大

unsloth-cli.py Ollama (#665) 2 weeks ago

README Apache-2.0 license

# unsloth

Start free finetune | Join our Discord | Buy Me a Coffee

**Finetune Llama 3, Mistral, Phi-3 & Gemma 2-5x faster with 80% less memory!**

## ✨ Finetune for Free

All notebooks are **beginner friendly**! Add your dataset, click "Run All", and you'll get a 2x faster finetuned model which can be exported to GGUF, Ollama, vLLM or uploaded to Hugging Face.

| Unsloth supports | Free Notebooks | Performance | Memory use |
|---|---|---|---|
| **Llama 3 (8B)** | Start for free | 2x faster | 60% less |
| **Mistral v0.3 (7B)** | Start for free | 2.2x faster | 73% less |
| **Gemma 2 (9B)** | Start for free | 2x faster | 63% less |
| **Phi-3 (mini)** | Start for free | 2x faster | 50% less |
| **Phi-3 (medium)** | Start for free | 2x faster | 50% less |
| **Ollama** | Start for free | 1.9x faster | 43% less |
| **ORPO** | Start for free | 1.9x faster | 43% less |
| **DPO Zephyr** | Start for free | 1.9x faster | 43% less |
| **TinyLlama** | Start for free | 3.9x faster | 74% less |

- **Kaggle Notebooks** for Llama 3 (8B), Gemma 2 (9B), Mistral (7B)
- Run Llama 3 conversational notebook and Mistral v0.3 ChatML
- This text completion notebook is for continued pretraining / raw text

Custom properties
12.2k stars
86 watching
795 forks
Report repository

**Releases** 1

2x faster Gemma 2 Latest
2 days ago

**Sponsor this project**

ko-fi.com/unsloth

**Packages**

No packages published

**Contributors** 21

+ 7 contributors

**Languages**

Python 100.0%

Alpaca + Llama-3 8b Unsloth 2x faster finetuning.ipynb

檔案 編輯 檢視畫面 插入 執行階段 工具 說明 無法儲存變更

共用 Gemini

+ 程式碼 + 文字 複製到雲端硬碟

T4 RAM 磁碟

To run this, press "*Runtime*" and press "*Run all*" on a **free** Tesla T4 Google Cc

unsloth | Join our Discord | Buy Me a Coffee Join Discord if you r

*us on Github* ⭐

To install Unsloth on your own computer, follow the installation instructions page here.

You will learn how to do data prep, how to train, how to run the model, & how Llama.cpp).

**[NEW] Llama-3 8b is trained on a crazy 15 trillion tokens! Llama-2 was 2 tril**

Use our Llama-3 8b Instruct notebook for conversational style finetunes.

```
[ ] %%capture
    # Installs Unsloth, Xformers (Flash Attention) and all other p
    !pip install "unsloth[colab-new] @ git+https://github.com/unslothai
    !pip install --no-deps xformers "trl<0.9.0" peft accelerate bits
```

- We support Llama, Mistral, Phi-3, Gemma, Yi, DeepSeek, Qwen, TinyLla Hermes etc
- We support 16bit LoRA or 4bit QLoRA. Both 2x faster.
- `max_seq_length` can be set to anything, since we do automatic RoPE Sc kaiokendev's method.
- With PR 26037, we support downloading 4bit models **4x faster!** Our re Mistral 4bit models.
- **[NEW]** We make Phi-3 Medium / Mini **2x faster!** See our Phi-3 Medium

```
[ ] from unsloth import FastLanguageModel
    import torch
    max_seq_length = 2048 # Choose any! We auto support RoPE Scal
    dtype = None # None for auto detection. Float16 for Tesla T4
    load_in_4bit = True # Use 4bit quantization to reduce memory
```

資源 FileNotFoundError

你

請說明錯誤:

```
FileNotFoundError: [Errno 2] No such file or direc
```

Gemini

## Hypothesis

The error message "FileNotFoundError: [Errno 2] No such file or directory: './content/drive/MyDrive/unsloth.Q4_K_M.gguf'" indicates that the Python code is trying to copy a file to a destination path in Google Drive that doesn't exist yet. The directory './content/drive/MyDrive/' likely exists (because you mounted your Google Drive), but the file 'unsloth.Q4_K_M.gguf' does not yet exist within that directory.

## Suggested Changes

```
import shutil
local = './model/unsloth.Q4_K_M.gguf'
google_drive_dir = './content/drive/MyDrive/' # Onl
filename = 'unsloth.Q4_K_M.gguf'
google_drive = os.path.join(google_drive_dir, filena
shutil.copy(local,google_drive)
```

在這裡輸入提示

0 / 400

回應內容可能不正確或令人反感，這類資訊並不代表 Google 的觀點。 瞭解詳情

已連線至「Python 3 Google Compute Engine 後端 (GPU)」

Edge

| | 1 ChatGPT | 2 Open Source Model | 3 API | 4 SGSGPT |
|---|---|---|---|---|
| 方案 | Teams | Ollama + Open WebUI | API + Open WebUI | SGS員工 |
| 費用(NT) | 23,334/年 | 50,000(單次) | 20,000(單次) | 總公司支付 |
| 功能 | ★★★ | ★★★ | ★★★ | ★ |
| 離線使用 | 🔴 | 🟢 | 🔴 | 🔴 |
| 隱私安全 | 🔴 | 🟢 | 🔴 | 🟠 |
| 優點 | 功能最新 | 管控與安全性最高 | 同 1 但費用更少 | 安全? |
| 缺點 | 費用高 | 依賴硬體、自行維運 | 自行維運 | 功能最少 |
| 帳號數量 | 2 | ∞ | ∞ | 依申請 |

# LLM Tools

https://lmstudio.ai/

https://useanything.com/

LM Studio

- Support Open Source Models
- Running model as API service
- Chat with Model
- Support OpenAI API
- Support Azure OpenAI Service

> At least 8G RAM for `7B` model,
> 16G for `13B`, and 32G for `33B`

Anything LLM

# Taiwan Llama 3

https://github.com/MiuLab/Taiwan-LLM

- 台大資工　林彥廷博士製作
- 基於 Meta Llama-3 模型
- 使用繁體中文語料庫 Fine-Tuning
- TMLU 測試 `70B` 優於 Claude3、GPT4o
- 於台灣本地知識、法律等領域表現優異
- 整體性能優於原始 Llama 3 模型
- 企業(單次)訓練成本仍然偏高

> find more LLMs on HuggingFace 🤗

# Homework

> 🚨自由練習，不強制
> ⚠️注意AI表現會與電腦硬體相關

- 前往以下網址註冊帳號 (需審核)
  http://10.213.125.140/
- 嘗試進行幾項測試
  - 與AI對談，詢問任何問題
  - 上傳檔案，討論檔案內容
  - 其他深度操作
- 安裝一套 LLM Tool，並嘗試使用
- 嘗試更多的開源模型

# 😀 Thank you !

feel free to ask if you have any other questions.


> **OAD** / brian_li / #1429
> brian.li@sgs.com