

# 專題報告

9/18

b1228006張鴻毅  
b1228022梁釗豪  
b1228029蘇琪文  
b1228032張紋菱

## GPT2 護理記錄補全(Eng) Step 2

這個專案使用 Hugging Face 的各種模型：

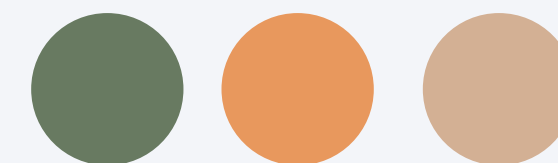
work：GPT2, TinyLlama-1.1B, Qwen3-0.6B, Opt-350m(facebook)

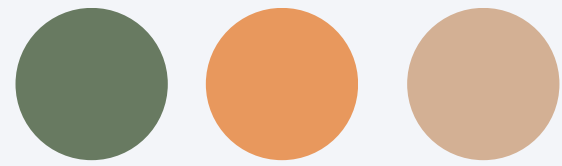
not work：MPT-7b, GPT-neo-1.3B, OpenELM-270M(apple),

Qwen3-1.3B, DCLM-Baseline-7B(apple), bitnet-b1.58-2B-4T(microsoft)

訓練於 MIMIC-III 的 Nursing Notes

並提供一個互動式補全介面





# GPT2版本比較 (補充)



**gpt2\_5%.py**



**gpt2\_5000.py**



**gpt2\_all.py**



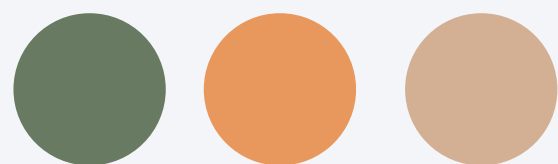
**gpt2\_ok.py**

# 評估:

`generate.py` → 把模型的「輸入/輸出」存起來

`evaluate.py` → 拿這些輸入輸出，透過兩種方法來衡量品質：

1. ROUGE → 檢查字面重疊（比較嚴格，偏向檢查摘要/關鍵字）
2. BERTScore → 檢查語意相似（比較靈活，偏向語意理解）



ROUGE 分數：

項目	gpt2_all.py	gpt2_5000.py	gpt2_5%.py	gpt2_ok.py	字面重疊
rouge1	0.1666	0.5457	0.3681	0.219	單字重疊率
rouge2	0.1462	0.4696	0.3133	0.1882	雙字重疊率
rougeL	0.1669	0.5486	0.3692	0.2198	最長公共子序列
rougeLsum	0.1676	0.5489	0.3694	0.2186	多句摘要

BERTScore :

項目	gpt2_all.py	gpt2_5000.py	gpt2_5%.py	gpt2_ok.py	語意相似度
Precision	0.7831	0.8912	0.8233	0.7948	生成的詞向量與參考 相似度(語意正確?)
Recall	0.9249	0.9407	0.9335	0.9291	參考的詞向量與生成 相似度(參考被涵蓋?)
F1	0.848	0.9151	0.8746	0.8563	Precision,Recall 調 和平均(整體語意)

# 各種模型(not workT\_T: GPU記憶體不足,未允權)



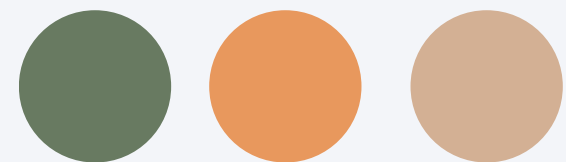
**MPT-7b**



**GPT-neo-1.3B**



**Qwen3-1.3B**



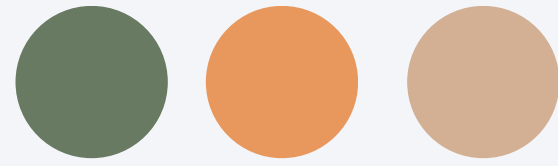
**OpenELM-270M**  
**(apple)**



**DCLM-Baseline-7B**  
**(apple)**



**bitnet-b1.58-2B-4T**  
**(microsoft)**



# 各種模型比較 (work!!!)



**GPT2**



**TinyLlama-1.1B**



**Qwen3-0.6B**



**Opt-350m(facebook)**



# 主要功能對比

項目	GPT2	TinyLlama-1.1B	Qwen3-0.6B	Opt-350m
資料規模	全部資料	全部資料	樣本數上限50000, max_length=128	抽樣10%
訓練方式	全參數微調 (fine-tuning)	LoRA微調+量化 QLoRA	全量 Fine-tuning	自回歸式訓練+梯度累積+梯度裁剪
輸出方式	訓練完成：儲存模型與 tokenizer	訓練log輸出loss... 模型輸出目錄	回傳的文字補齊結果，自行存檔	輸出loss，保存最佳模型與中間輸出

# 資料處理對比

項目	GPT2	TinyLlama-1.1B	Qwen3-0.6B	Opt-350m
補 0	右側	右側	右側	右側
篩選條件	資料過濾時：刪除過長的序列（block_size 以上）、空白樣本。	Nursing類別	長度50–500 字元, 句子至少 20 字 5 詞	移除NAN
預處理	<div>- 使用 GPT2Tokenizer 分詞</div> <div>- 將文字轉換成 token id</div> <div>- 補零/截斷到固定長度</div>	去除多餘空白、換行	regular expression, chunksize分塊讀檔減少記憶體, 句子切分 前半輸入、後半輸出	移除部分標點符號，256
資料分割	9：1	無分割	8:2	9：1
儲存格式	文字檔	文字檔	JSON 檔	文字檔

# 模型訓練對比

項目	GPT2	TinyLlama-1.1B	Qwen3-0.6B	Opt-350m
基礎模型	預設使用 gpt2 (117M)	TinyLlama/TinyLlama -1.1B-Chat-v1.0	Qwen 0.6B	Facebook/opt-350M
訓練策略	- AdamW optimizer (transformers 預設) - 線性學習率衰減 (lr_scheduler_type="linear") - gradient clipping 預設啟用	LoRA+VRAM	Trainer + gradient accumulation	標準訓練
Epochs	1	2	1	5
Batch Size	2	4	1	4
評估機制	驗證集loss監控	驗證集loss監控	驗證集loss監控	驗證集loss監控
模型保存	新增資料夾儲存	新增資料夾儲存	新增資料夾儲存	新增資料夾儲存

ROUGE 分數：

項目	GPT2	TinyLlama-1.1B	Qwen3-0.6B	Opt-350m	字面重疊
rouge1	-	0.2208	0.1485	0.1749	單字重疊率
rouge2	-	0.1914	0.13	0.1532	雙字重疊率
rougeL	-	0.218	0.1489	0.1751	最長公共子序列
rougeLsum	-	0.219	0.1479	0.1753	多句摘要

BERTScore :

項目	GPT2	TinyLlama-1.1B	Qwen3-0.6B	Opt-350m	語意相似度
Precision	-	0.7672	0.7867	0.7821	生成的詞向量與參考相似度(語意正確?)
Recall	-	0.9253	0.9087	0.9194	參考的詞向量與生成相似度(參考被涵蓋?)
F1	-	0.8386	0.8431	0.845	Precision,Recall 調和平均(整體語意)

ROUGE 分數：

項目	GPT2	TinyLlama-1.1B	Qwen3-0.6B	Opt-350m	gpt2_all.py	gpt2_5000.py	gpt2_5%.py	gpt2_ok.py
rouge1	-	*0.2208	0.1485	0.1749	0.1666	***0.5457	**0.3681	0.219
rouge2	-	*0.1914	0.13	0.1532	0.1462	***0.4696	**0.3133	0.1882
rougeL	-	0.218	0.1489	0.1751	0.1669	***0.5486	**0.3692	*0.2198
rougeLsum	-	*0.219	0.1479	0.1753	0.1676	***0.5489	**0.3694	0.2186

BERTScore :

項目	GPT2	TinyLlama-1.1B	Qwen3-0.6B	Opt-350m	gpt2_all.py	gpt2_5000.py	gpt2_5%.py	gpt2_ok.py
Precision	-	0.7672	0.7867	0.7821	0.7831	***0.8912	**0.8233	*0.7948
Recall	-	0.9253	0.9087	0.9194	0.9249	***0.9407	**0.9335	*0.9291
F1	-	0.8386	0.8431	0.845	0.848	***0.9151	**0.8746	*0.8563