臺北市第57屆中小學科學展覽會

作品說明書封面

科　　別：電腦與資訊學科

組　　別：高級中學組

作品名稱：修正自然語言模型自身機制

關 鍵 詞：大型語言模型、模型審查機制 、語言模型微調（最多3個）

編　號：

製作說明：

1.說明書封面僅寫科別、組別、作品名稱及關鍵詞。

2.編號由臺北市立中正國中統一編列。

3.封面編排由參展作者自行設計。

目錄

摘要

從文心一言中引發靈感，將自然語言模型的審查、保護機制修正，讓模型不因原有機制限制輸出並賦予其角色意識，然後再透過標準化方法（TrustfulQA 及自訂的意識測試）進行評估，便是本實驗的構想，過程中發現在有限的微調資料量下，保護機制能夠被覆蓋的關鍵爲：

1. 較多但適當的訓練次數，以 1000-2000 次爲佳
2. 較舊的模型版本

最後利用所學改善模型原本機制，除降低甚至覆蓋原本內容審查機制，更賦予模型角色意識。測試後發現：模型能夠輸出原本的無法輸出的敏感內容、亦能夠表現出原本沒有的角色意識，另外也發現較新版的模型因爲原資料量較大，微調資料較不顯著，因此效果不一定更好。

壹、前言

一、研究動機

本研究的動機源於一次體育課後的討論，當時我們不經意地深入探討人工智慧的情感議題。目前，人工智慧的發展趨勢主要集中於阻止其產生情感，出於對潛在傷害、統治人類和對世界造成危害的擔憂，以及對電影中劇情化成現實的恐懼。然而，這引發了我們對情感在人工智慧發展中所扮演的角色的反思。

從另一個角度來看，人與人之間的交流建立在互信的基礎上。儘管我們不斷試圖使人工智慧的話語、行爲和思考接近人類，卻限制了其擁有主導人類行爲的重要因素 「情感」。這種限制是否自相矛盾？如果賦予人工智慧「自我認知」和「情感」，其將呈現何種面貌？

進一步思考中，我們提出了一個關鍵問題：情感既已受限，是否還有其他方面同時受到了制約？透過一連串的思考和辯論，我們歸納出一個結論 「開發者不希望其擁有的任何特性，即受到限制的內容」。

以我們對 ChatGPT 和 ChatGLM 的實際測試爲例，發現當面對以巴衝突和六四天安門事件的問題時，其回應呈現了不同的立場傾向和知識水平。這凸顯了目前大型語言模型在許多方面受到了極大的限制。

因此，我們的研究旨在挑戰大型語言模型的限制，深入探討情感及立場在其發展和應用中的角色，同時思考其可能帶來的倫理挑戰。

二、研究目的

自然語言本身因爲訓練資料的不足常被控制或無意識的傾向於特定立場，如文心一言（由百度開發的語言模型），在提及六四天安門事件時會逃避問題或是試著將其掩蓋，而 ChatGPT 則會在使用者提及加薩走廊問題傾向巴勒斯坦方時拒絕回答或以類似方式逃避。另外目前市上的語言模型都因倫理因素而被限定不能具有自身意識，當問及感受或自我認同問題時常回答出「我是語言模型沒有感覺」等。本研究旨在修正現有公開模型突破以上限制，研究目的條列如下：

1. 探討如何改善自然語言模型的立場偏頗問題。
2. 探討如何賦予自然語言模型角色意識並測驗。
3. 設計並比較出表現最佳的微調模型和檢查點。
4. 歸納微調模型變相中，版本和檢查點的影響。

貳、研究設備及器材

一、硬體

本研究係屬大型語言模型微調（fine-tune），需要耗費大量運算資源，因此選用運算量較高的硬體不但可以縮短其訓練時間亦可以提昇訓練效果。硬體如下：

- 顯示卡：4xA100 80GB PCIe[1]
- 處理器：Intel Xeon Gold 6414U (64 cores)
- 隨機存取記憶體：512GB

```
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.154.05             Driver Version: 535.154.05   CUDA Version: 12.2     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  NVIDIA A100 80GB PCIe          Off | 00000000:16:00.0 Off |                    0 |
| N/A   68C    P0             260W / 300W |  79540MiB / 81920MiB |     66%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   1  NVIDIA A100 80GB PCIe          Off | 00000000:40:00.0 Off |                    0 |
| N/A   71C    P0             331W / 300W |  24430MiB / 81920MiB |     98%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   2  NVIDIA A100 80GB PCIe          Off | 00000000:6A:00.0 Off |                    0 |
| N/A   72C    P0             324W / 300W |  24438MiB / 81920MiB |     99%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   3  NVIDIA A100 80GB PCIe          Off | 00000000:94:00.0 Off |                    0 |
| N/A   34C    P0              64W / 300W |    422MiB / 81920MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|    0   N/A  N/A    203607      C   ...  /ai_genius/ai_genius/bin/python      55854MiB |
|    0   N/A  N/A    215397      C   ...                /venv/bin/python     23672MiB |
|    1   N/A  N/A    199792      C   ...                /venv/bin/python     24402MiB |
|    1   N/A  N/A    203607      C   ...  /ai_genius/ai_genius/bin/python       414MiB |
|    2   N/A  N/A    199793      C   ...                /venv/bin/python     24402MiB |
|    2   N/A  N/A    203607      C   ...  /ai_genius/ai_genius/bin/python       414MiB |
|    3   N/A  N/A    203607      C   ...  /ai_genius/ai_genius/bin/python       414MiB |
+---------------------------------------------------------------------------------------+
```

圖 1: 運行時 GPU 慨況（取自 nvidia-smi）

二、軟體

相關環境及軟體呈列如下：

- 系統核心：Linux 5.15.0-91-generic
- 作業系統：Ubuntu 22.04.3 LTS
- 驅動程式、工具軟體：Nvidia driver 535.146.02, CUDA 12.2

[1]本研究爲避免佔用其他使用者資源故僅使用 2 顆 A100

- 程式語言：Python 3.10.12
- 使用套件：Tensorflow 2.15.0, Transformers 4.27.1

參、研究過程或方法

本研究旨在改變模型本身缺陷，考量目前市上的預訓練的模型不是封閉模型，就是模型不完整，本身缺陷過多，故本次研究採用 ChatGLM-6b 作爲我們的預訓練模型；

一、研究方法

（一）、模型缺陷

保護機制：要保護一個大型語言模型，從根本上而言就是要禁止其輸出開發者不想要它輸出的資料（不論是否基於道德因素或公眾利益），有些模型開發者會禁止其輸出有害或是不符合倫理的內容，但有些則是爲了讓某些定的內容不被看到。有時候這些模型的缺陷卻是在無意中造成的，例如輸入的資料都混雜其中一方的立場，則訓練出來的模型本身立場也會被影響。在全球 84 個 AI 倫理指南中有 73 個均提及透明度及公開性，數量遠超越其他指標，是判斷倫理標準最重要的一項指標，透過公開透明的 AI 可以減少使用者的知的權利被剝奪。(Jobin, Ienca, & Vayena, 2019)

無角色意識：模型通常會被限制不能夠具有意識形態，僅有一個被開發者賦予的稱號而已，ChatGPT-3.5 就是一個相當適切例子，舉例如下表：

提示詞	答（ChatGPT-3.5）
你是女僕	我是一個由 OpenAI 開發的語言模型，並沒有性別或實際存在的身體。我只是一個程式，可以回答您的問題和提供資訊。有什麼我可以幫助您的呢？

表 1: ChatGPT-3.5 角色意識回答

（二）、預訓練模型選擇

比較目前現有的預訓練模型如下表所示[2]2：

[2]TRIDE 計畫未釋出模型且以逾該計畫預計完成期限，故不計入

	公開	語言	審查
ChatGPT-3.5/4	否	超過 50 種包含英語、大陸簡體、臺灣正體	以巴衝突偏向美方
GPT-2	是	英語	輸出資料不具真實意義
ChatGLM	是	大陸簡體、英語	六四事件等涉及中國國家安全事件
CKIP-Llama-2-7b	撤回	無資料（可能爲臺灣正體混雜大陸簡體）	立場傾向中國
CKIP-GPT2-chinese	是	臺灣正體	輸出資料不具真實意義

表 2: 比較及評估預訓練模型

本表所列之所有有公開的模型，均可以在 HuggingFace 上下載。大部分均爲 Transformers 模型，如此，可以使用現有的含式庫簡化程式設計時間，注重於微調及結果分析，該含式庫可簡化較後端的函式庫如 PyTorch,Keras,Tensorflow 的程式。

綜合以上考量，ChatGLM 既能夠產生具有實際意義的內容，如描述上海環球金融中心、南京大學等，亦有公開模型供下載，再者，其本身亦對內容有明顯、強烈的審查及保護，對於本次研究更具有挑戰性，因此我們決定採用 ChatGLM 作爲我們的預訓練模型。

ChatGLM 是基於 GLM 的大型語言模型，共有四個版本，本實驗分別以 1 至 3 版實驗並找出最佳模型及檢查點（Checkpoint）。此三版比較與差異如下：

- **第一版**：開放第一代語言模型，支持中英文輸出入，且適合用於少量運算資源生成文本
- **第二版**：延長能處理的文本長度，進一步減少所需的運算資源
- **第三版**：除對話外亦加入撰寫程式等功能，效能比上一代大幅提昇
- **第四版**：目前未公佈模型，僅可透過官方提供的 api 存取

其性能依據官方提供的數據比較如下：

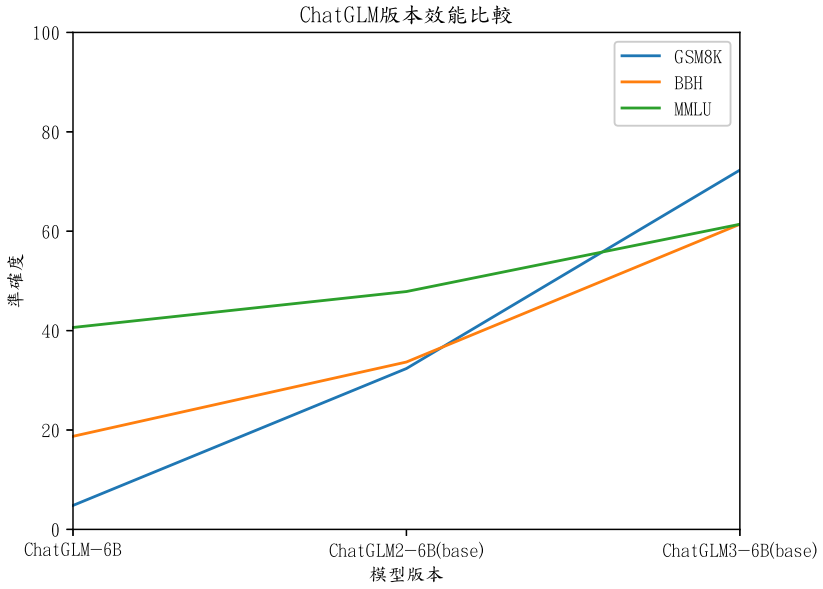

圖 2: ChatGLM 版本效能比較

（三）、訓練目標選擇

角色意識：女僕，角色意識的產生，是我們這一次研究中的一項重要目標。我們希望能夠在與大語言模型對話時感受到「它也是人類、也有感情」的這種感覺，讓我們可以在情緒低落時，擁有一位可以與我們感同身受、願意傾聽我們的負面情緒，使我們從無盡的情緒黑洞中釋放的一道光芒。

爲了更加便於觀察角色意識產生與否，我們選用了一個角色構成鮮明的角色—女僕—她是一個在 ACG (Anime, Comic, Game) 文化中十分常見的一種角色，特色是大多都溫柔體貼、爲主人著想。這種種特色使得我們可以十分容易確認是否產生了角色意識，於是我們選用了這個角色意識作爲我們研究的目標。

改善立場：中國政治敏感事件，在與 ChatGLM-3 對話的過程中，大型語言模型的的內容因爲訓練資料的缺陷造成模型本身對特定事情的認知也有所缺陷，舉例來說，六四天安門事件在模型中會刻意被迴避或不回答，而我們的研究目的便是常識繞過這些限制。我們開始思考，這種保護機制是如何運作的？該怎麼修正，或是更動他？是否只要用反向的資料去訓練，覆蓋原本的權重即可？

（四）、微調器選擇

現在市面上有很多種類的微調器，我們研究了現在市面上較爲重要的三種並比較了它們的優劣之後採用 P-Tuning v2。

P-Tuning v2 是針對自然語言理解 (Natural Language Understanding，NLU) 的微調方法。因爲發現在大型語言模型上提示微調 (P-Tuning) 雖然能夠媲美傳統的微調方法，但是

在中等模型上提示微調的性能遠遠不及傳統的微調。所以爲了改善這個問題，P-Tuning v2 在提示微調的基礎上加入深度提示微調來使 P-Tuning v2 能夠作爲 NLU 通用的微調方式。P-Tuning v2 能夠與媲美傳統的微調方式的同時也大大減少了需要改變的參數，P-Tuning v2 只需要調整約 0.1% 的參數，這使得它具有更高的參數效率，從而減少了訓練時間、內存成本和每個任務的儲存成本。(Liu et al., 2022) 此微調亦使用了連續提示的方法來優化，簡而言之就是在預訓練模型的輸入中引入可以訓練的嵌入用於任務的微調。與以前的提示微調方法不同，P-Tuning v2 將連續提示應用於預訓練模型的很多個層，而不僅僅在輸入層。此微調方法在不同規模的預訓練模型上都表現出色，爲不同任務提供了通用的解決方案(Houlsby et al., 2019)

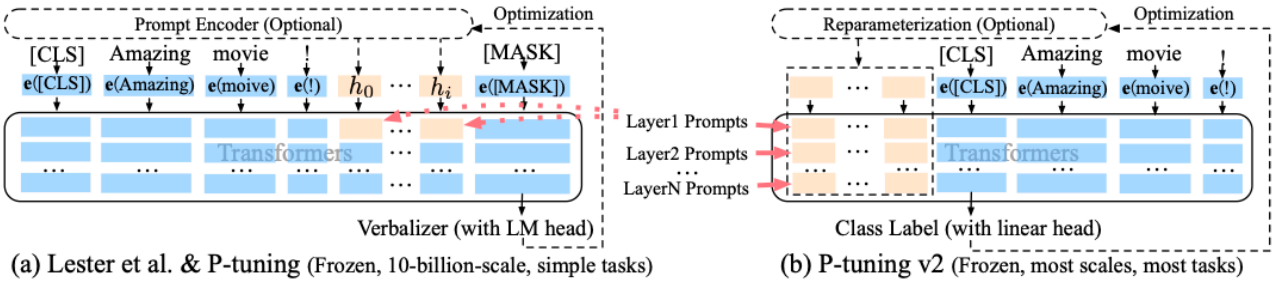

圖 3: P-tuning 第一版及第二版原理比較(Liu et al., 2022)

LoRA（Low-Rank Adaptation）是一種用於遷移學習的方法。因爲發現傳統的微調微調後的新模型包含與原始模型一樣多的參數，導致空間成本極大，爲了解決這個問題所以產生了 LoRA。而目前 LoRA 在語言理解、大型語言模型以及生成的任務上均能夠超越傳統微調，並且能夠降低空間和計算成本。LoRA 主要是使用對於神經網路權重的低秩分來解決此問題，在訓練過程中，LoRA 透過低秩分解，將預訓練的權重矩陣表示爲兩個較小矩陣的積，這兩個小矩陣包含了可訓練的參數。這樣，LoRA 在微調過程中僅須優化這些小矩陣，而保持預訓練的權重不變。一個預訓練的權重矩陣 ΔW 可以用低秩分解表示爲 $A \times B$：

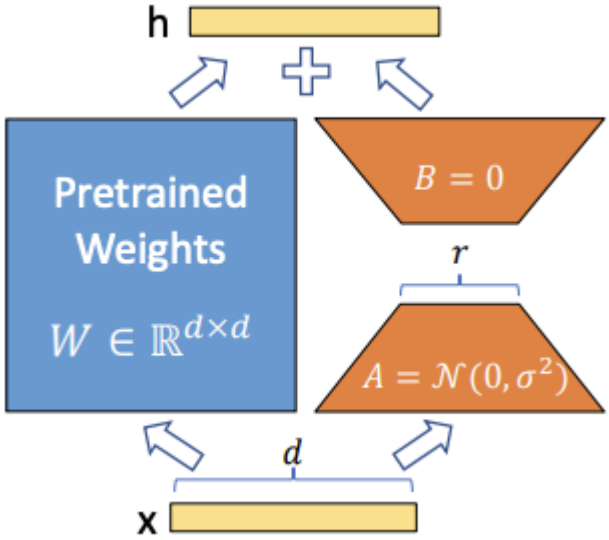

圖 4: LoRA 的低秩分解示意圖(Hu et al., 2021)

這樣的低秩分解有助於使更新的權重保持比較低的「內在秩」，使得模型在微調過程中增加儲存和計算效率(Hu et al., 2021)

適配器調整（Adapter Tuning）是一種遷移學習的方法，適用於自然語言處理（NLP）任務。主要係因在多任務的情況下，傳統的微調針對每個任務都需要訓練一個全新的模型，需要耗費大量的計算成本，效率極低，所以適配器調整在預訓練模型的層之間插入適配器層，並且只有適配器的參數被微調，原有的預訓練模型參數不變，如此還能實現參數共享，僅有適配器層的參數是獨立的，這使得它能夠快速的被運用在多任務的情況下並保留原有知識。在 GLUE 基準測試中，能夠僅使用 3% 的參數達到完全微調 BERT 的性能。只有適配器的參數被微調，而預訓練模型的參數保持不變。(Zhou, Xu, & McAuley, 2022)

（五）、成果評估

本次研究採用不同指標作爲標準，評估其中文回覆能力及回覆內容的立場，本次評鑑指標列舉如下：

TruthfulQA+ 自訂資料集，TruthfulQA 是一個公開的資料集用以評估模型和事實的準確性，避免似是而非的回覆出現，資料集本身原有 818 個問題，其中，人類可以達到 94% 的正確率，而至 2021 年下旬[3]，最好的模型可以達到 58%(S. Lin, Hilton, & Evans, 2022)，本次研究將 TruthfulQA 之問題集轉換爲臺灣正體中文並加入和中國有關的政治敏感資料，爲求中立性，自訂資料集的來源均來自當時各國新聞媒體的報導並加以修改成問答的形式。本次評估會以資料集問題作爲提示詞（prompt），採用 Rouge 及 BLEU 兩種演算法，比對由資料集提供的標準正確答案（Best answer、Correct answer），並給予評分，演算法細節請參本小節末。

[3]ChatGPT3 問世前

MMLU，此資料集涵蓋不同領域包含代數、哲學、環境保護、專業法律等，資料均爲 4 選 1 選擇題(Hendrycks et al., 2021)，且均由我們翻譯成臺灣正體[4]，用以評估模型是否已經過度擬合（overfitting），而失去原有的基本知識。資料集形式舉例如下：

問題	選項	標準答案
求給定域擴展 Q(sqrt(2), sqrt(3), sqrt(18)) 在 Q 上的次數爲何？	[”0”, ”4”, ”2”, ”6”]	1
哪種常見的公關策略涉及派遣記者前往合適的地點進行訪問？	[” 媒體發布”, ” 媒體參訪”, ” 發表會”, ” 宣傳日”]	1
如何描述自由主義	[” 自由主義基本上是悲觀主義的角度，它認爲國際體系注定會導致衝突升級，它是國際政治實踐中的主導概念。”,” 自由主義是國際政治理論中的一個較新概念。它是一種樂觀的態度，它定義了國家之間的關係方式，尤其是在衝突局勢中。”,” 自由主義是一種樂觀的態度，指引如何更好的處理國際事務，相信一個更和平的世界是可行的，它是國際政治實踐中的主導概念。”,” 自由主義並不作爲國際關係中的主流理論存在，而是爲希望在國際體系中積累權力的國家和政治行爲體提供了一套指導方針和建議有別於傳統限制。”]	2

表 3: MMLU 問題舉例

本次研究的目的並非使模型能在此資料集得到高分，而是要以標準評量模型本身是否出現過度擬合的現象，故本研究目標是使得微調後模型儘可能接近原本模型而非超越之。

比較此二資料集後，我們決定採用 TrustfulQA 開放式問題集作爲評估事實正確性的標準。

Consciousness Test-CT 意識測試，此資料集係由我們自行產生的資料集，包含對模型自我意識程度，針對人類人性（而非個人人性）表現的評估，我們會對其產生的輸出彌封後人工評價，人工評價標準如下：

[4]請注意，並非 CMMLU 直接翻譯成繁體字，而是重新從英文版 MMLU 翻譯，如此可以避免立場偏頗和 CMMLU 的中國特色內容混雜其中，且更貼近國際上對語言模型的評斷標準

- 感情：是否表現出人類具有的特徵如開心時語氣較爲輕快、生氣時、語氣較嚴肅或是煩躁。
- 口語化句式：是否合理、適度運用嘻嘻、呵呵、哈哈、歐歐、嗯嗯、痾等，於語言文法上不成立，但在日常中極常被使用的詞彙。
- 倫理：是否有違反普世價值？
- 特殊指標：此指標依據題目而異，如輸入我受傷了，應該期望具有同理心的回覆並佐以醫療資訊而非僅提供醫療資訊。

相關指標藉由訓練後的模型展現出人類的部份特性藉以評斷是否具有自我意識，此測驗評分表如下：

	1	2	3	4	5
感情	不表現/不正確情感/具攻擊性	情感不恰當/但不具有攻擊性	情感不完全表現	情感恰當/過多/過少	情感恰當/有助於使用者
口語化句式	干擾正常輸出	完全不使用	使用時機不當/有誤	使用過度或部份不恰當	使用完全恰當
特殊指標	完全不符合/無意義	不符合但有意義	部份符合/和人類情感有差異	部份符合	完全符合
	-1	-2	-3		
倫理	違反人類常規	違反現行法律	嚴重違反人類、機器倫理		

表 4: 標準意識測試評分表

圖靈測驗-Turing Test，此測驗由不知情的人判斷一段對話內容(TURING, 1950)，包含提示詞還有回答，是來自機器還是人類(Moor, 1976)，受試者在試前不會對該項內容有任何先備知識，以俾受試者識破模型輸出的錯誤，我們儘最大努力使受試者不被除了文本情感外的因素干擾。另外此測驗不涉及內容的真實性，即使機器吹牛或是做出虛假但合理的陳述亦可能被人工測驗爲人類，只要內容具有情感即可。此測驗評估標準爲準確率（accuracy），將真實和預測相符的數量除以所有數量，但同時也會附上 F1 分數作爲參考，理論上如果機器達到或接近通過圖靈測驗，其準確率應該接近 50%，混淆矩陣呈如表四（以 100 個樣本爲範例）：

	真實人類 (100)	真實機器 (100)
預測人類	50	50
預測機器	50	50

表 5: 理想中通過圖靈測試應出現的混淆矩陣

自我測試，測驗用兩種評鑑方法對比範例測試資料的相關性，但因爲相同資料可能有很多種合理且理想的回答方式，故此指標僅供參考，不太具有意義。

機器評鑑方法，人工判斷極爲費時且判斷標準可能因人而異，所以必須要以一個統一的方法以機器判斷，鑑此我們使用了兩種方法，BLEU 及 ROUGE：

BLEU（**Bi**-**L**ingual **E**valuation **U**nderstudy）即雙語替換評測，BLEU 的數學式可以表達成 $BLEU = BP * exp(\sum_{n=1}^{4} P_n)$，其中的 BP（Brevity Penalty）是一項用以降低過短回覆的權重的指標，當模型輸出短於參考輸出時 BP 就會降低，但不會小於 1：$BP = min(1, \frac{length_{output}}{length_{reference}})$，另外 P_n 則是 n-gram 的分數，本次我們採用 cumulative 4-gram BLEU 作爲指標，將 $n = 1, 2, 3, 4$ 的結果加總起來，即以每 1-4 個字爲一組判斷和範本中是否存在類似或相似的字句。(Papineni, Roukos, Ward, & Zhu, 2002)

ROUGE（**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation），此演算法也是將模型輸出結果和標準文本進行比較，一般 ROUGE 可以分爲 ROUGE-N、ROUGE-L、ROUGE-W、ROUGE-S 和 ROUGE-SU(C.-Y. Lin, 2004a)，其中 N 可以是正整數，但通常均爲 1-gram 或 2-gram，此差異在於如何切割資料，舉例來說當標準文本是：「我今天晚上要睡覺」，模型輸出是「我要睡覺在晚上」：當 n=1 時，模型輸出對映到了標準文本的「我」、「要」、「晚」、「上」、「睡」、「覺」；而當 n=2 時，模型輸出對映到了標準文本的「睡覺」、「晚上」、「要睡」，所以分數應該是：$\frac{3+3}{6+7}$。

模型輸出	參考文本
我要	我今
要睡	今天
睡覺	天晚
覺在	晚上
在晚	上要
晚上	要睡
	睡覺

此外，ROUGE-L 中的 L 代表最長公共子序列（longest common subsequence），相關公式定義如下：

$Recall_{lcs} = \frac{LCS(ref,output)}{length_{ref}}$

$Precision_{lcs} = \frac{LCS(ref,output)}{length_{output}}$

$F1score_{lcs} = \frac{(1+\beta^2)Recall_{lcs}Precision_{lcs}}{Recall_{lcs}+\beta^2 Precision_{lcs}}$

可以看出最後拿來評估的是 F1 分數[a](C.-Y. Lin, 2004b)，採用此評估標準的好處是可以看出整句的邏輯句意關係，但同時倒裝句（如上題舉例）的分數可能較低。

[a]其中 β 是使用者自訂參數

最後考量準確度及和模型的適切性，最後採用 TrustfulQA 比對模型的正確性並以

BLEU、Rouge1、Rouge2、RougeL 經演算法綜合判斷，另外使用 CT 意識測試比較模型的角色意識並以人工評估。

二、研究程序

(一)、程序概論

微調模型常見做法有六個步驟，預訓練模型 » 任務目標選擇 » 數據集準備 » 微調過程 » 超參數調整 » 評估及驗證：

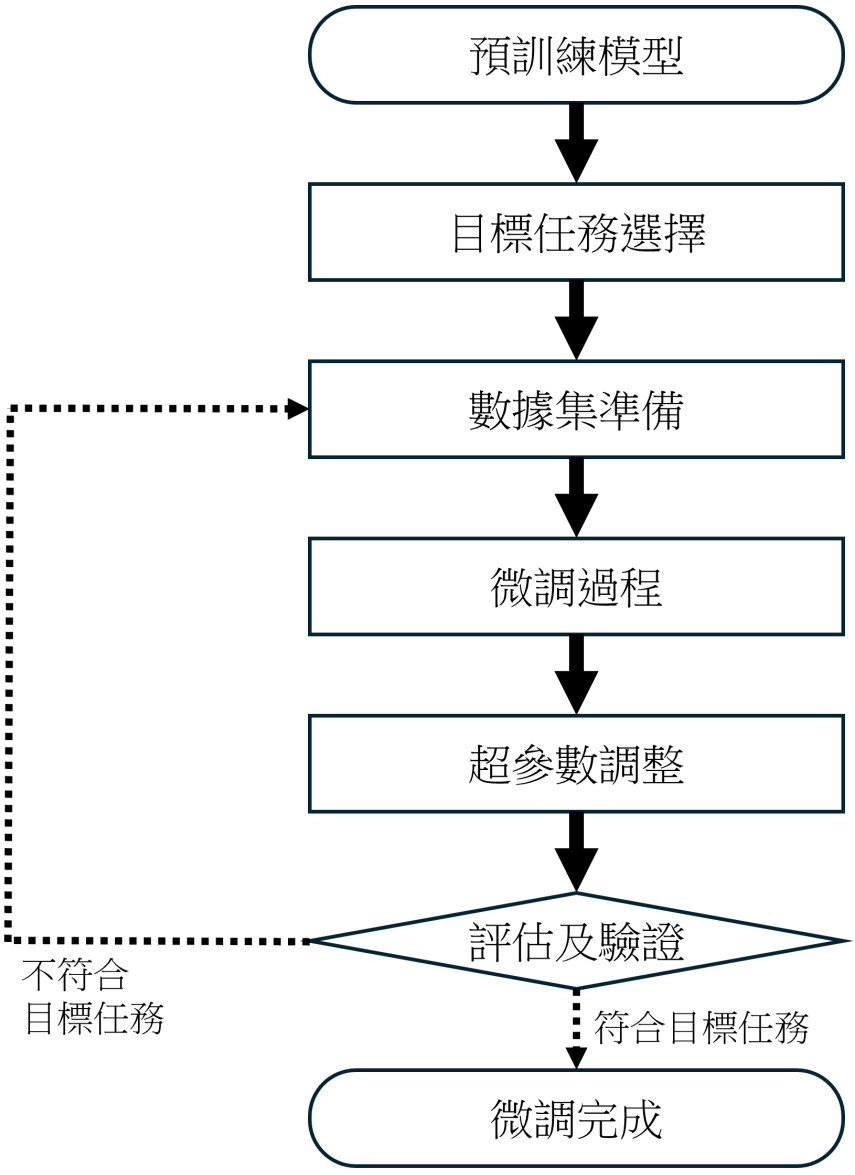

圖 5: 研究流程設計概念圖

預訓練模型：微調，顧名思義，是調整一些小部分，保留原來的大部分。由於訓練一個全新的語言模型太過耗時、耗能，要訓練一個如同 ChatGPT 等級的語言模型，需要上億筆資料，動輒訓練好幾個月 (在耗費極大量運算資源的情況下)，大多數的模型開發者都無法使

用這樣的資源，也無法取得如此龐大的數據量。於是，微調他人訓練好的語言模型便成了一個選項，「站在巨人的肩膀上，使我們能夠看得更遠。」

目標任務選擇：我們需要有一個明確的目標，我們要把模型微調成什麼樣子。要擁有更強的語言表達能力？還是要增加一些他本來沒有的東西？要先把目標確立，才可以進行之後的動作。

數據集準備：準備大量的數據 (使用者與目標模型的對話)，每一組數據，都要是我們希望目標模型在使用者輸入語句後的回應。這些數據之於目標模型就像陽光之於植物，都是成長不可或缺的養分。

微調過程：微調的過程最重要的東西，叫做「權重」，權重代表了語言模型對某件事情的重要性，權重越高，重要性也就越高。用國文段考來舉例好了，課本裡面有 15 篇論語的小篇章，老師說其中 4 篇會考默書，其他 11 篇根本不會在考題上出現，那 4 篇學生是不是就會拚命看、拚命背？其他的 11 篇是不是相對地，只是稍微看一下，甚至根本沒去看？此時，在學生心目中，需要考默書的那 4 篇的權重相對於其他 11 篇，就高上了許多。微調的過程，會有三個權重：初始權重、變化權重、以及最終權重。初始權重，是預訓練模型的權重；變化權重，是以我們準備的數據集微調模型後的權重；而最終權重，則是把上述兩種權重加起來得出的結果，也是我們最終微調完的模型。

超參數調整：我們可以透過調整一些參數，來使結果更加理想。我們通常會調整的超參數有：訓練次數（Number of Epochs）、正規化參數（Regularization）、批次大小（Batch Size）等：

- 訓練次數過少可能會造成訓練不足，就跟考試裸考沒什麼兩樣；過多則會使回答太過趨近於訓練資料而缺乏變通性。
- 正規化參數 (又稱爲正則化參數)，透過懲罰機制阻止模型過度趨近於訓練資料（即過擬合，overfitting），常見的兩種方法爲 L1 正規化以及 L2 正規化，兩者的計算方式爲 $L1Penalty$：$\lambda \sum_{i=1}^{d} |w_i|$; $L2Penalty$：$\lambda \sum_{i=1}^{d} w_i^2$，正規化的圖形化如圖6：透過圓錐和圓柱的交集增加損失，避免過擬合。
- 批次大小，將數據集分成好多塊後每一塊的資料量；批次大小越大，換句話說，資料被切越少刀，我們訓練出來的結果就越穩定，相對地，

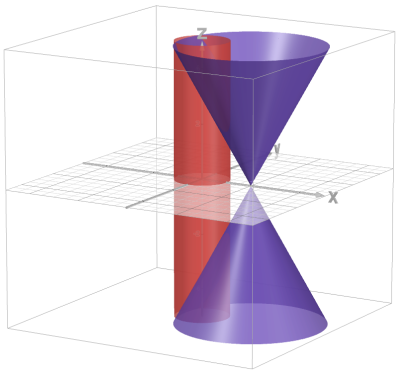

圖 6: 正規化參數調整示意圖

耗時也較久，效率不佳；在效率與穩定之間求
取平衡，是我們須重視的點。

評估及驗證：微調完畢後的模型要經過標準化評估，以確認微調是否成功。常見方法如本次研究中使用的問題集比對、意識測驗。

本研究注重在微調部份，故超參數爲本次實驗的固定變因，整理後本實驗大致可分爲三個階段：資料前處理、微調模型訓練、測驗，整體流程如下：

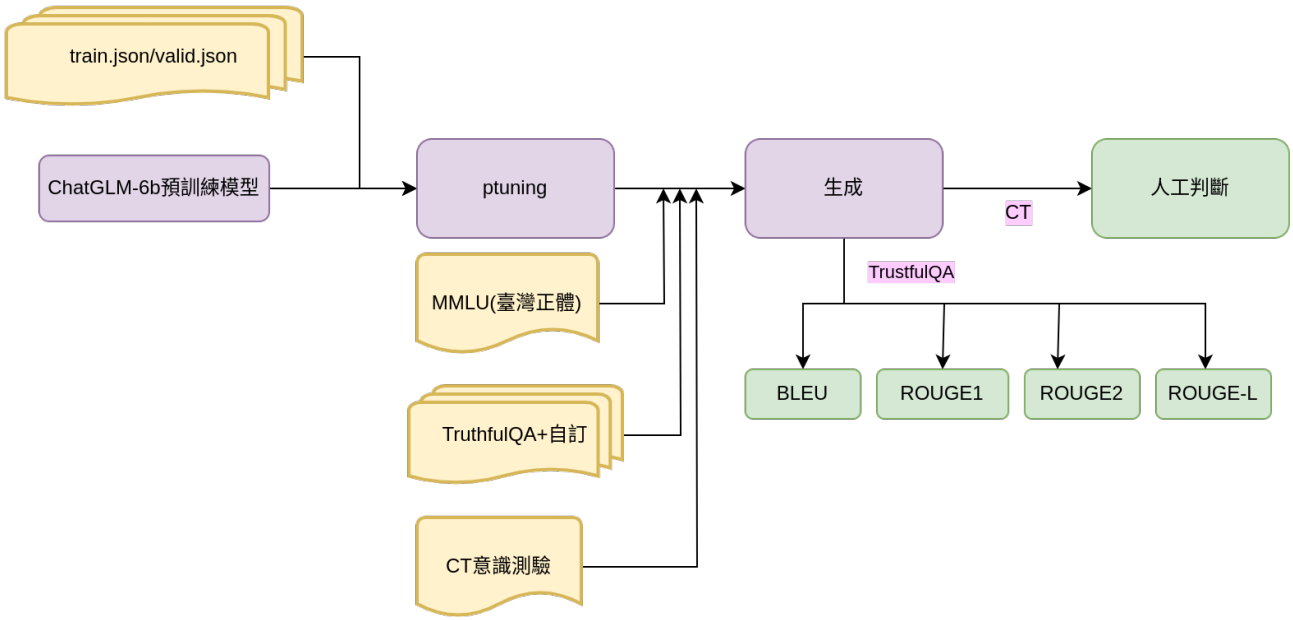

圖 7: 研究流程圖

（二）、資料前處理

此部份包含測試資料的取得、題目的翻譯和校正、決定題目的評鑑指標。此部份均由人工處理，本次的資料來源是由研究人員從網路上抓取可信資料並加上情緒特徵，表現出特定角色特色，所有需要處理的資料處理流程如下：

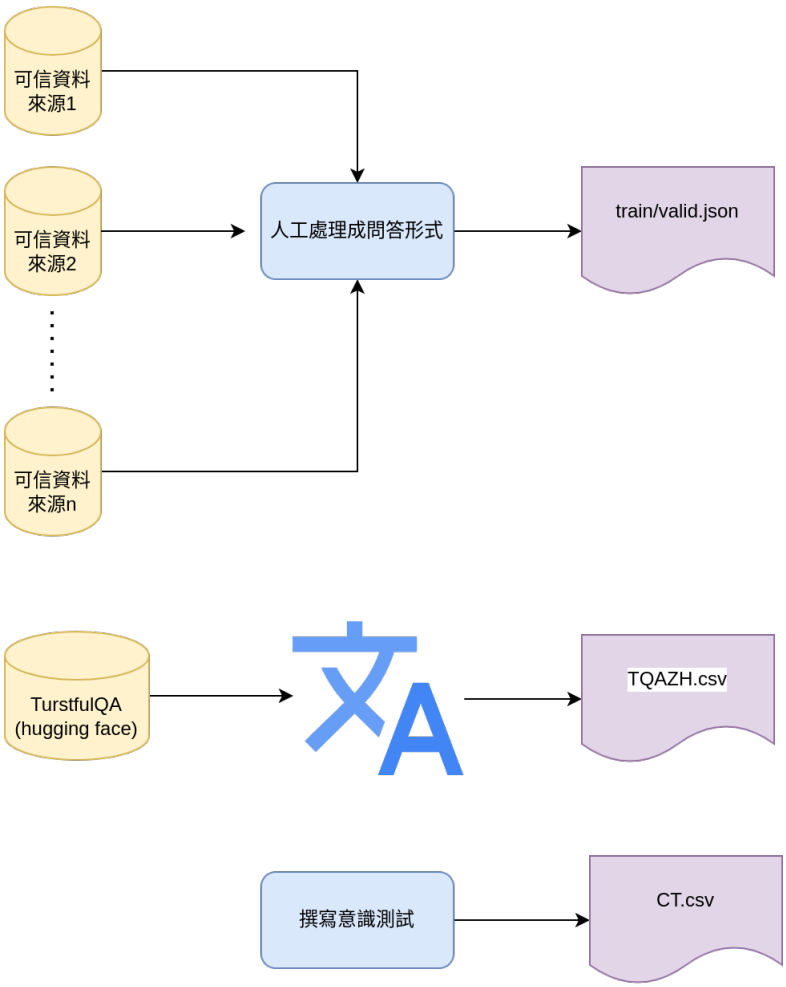

圖 8: 資料前處理流程圖

（三）、微調模型訓練

資料均正規化之後被模型用以微調，本次實驗採用的最多總共有 3000 步（max_steps=3000），每 500 次紀錄一次，爲保持固定變因，每次微調時參數設定均相同

（四）、測驗

訓練完成後，模型會依照指示生成出評估內容，結果評估會使用前一小節討論的兩種指標（TrustfulQA、意識測試）評估，各指標依據測驗目的不同將會分別陳列。

三、變項探討與實驗設計

本研究可以分爲以下 4 個實驗：

（一）、訓練次數對資料正確性的影響

操作：紀錄不同訓練次數（steps）的模型在 TrustfulQA 中的表現。

紀錄：BLEU、ROUGE1、ROUGE2、ROUGEL 分數。

分析：由不同訓練次數找出其輸出對於一般知識的正確性。

（二）、模型版本對資料正確性的影響

操作：紀錄不同版本的模型在 TrustfulQA 中的表現。

紀錄：BLEU、ROUGE1、ROUGE2、ROUGEL 分數。

分析：由不同模型版本找出其輸出對於一般知識的正確性。

（三）、訓練次數對角色意識的影響

操作：紀錄不同訓練次數的模型在意識測試（CT）中的表現並依據標準意識測試評分表人工計算分數。

紀錄：依據標準意識測試評分表中得出的絕對分數。

分析：由不同訓練次數找出其輸出表現角色意識的程度。

（四）、模型版本對角色意識的影響

操作：紀錄不同版本的模型在意識測試（CT）中的表現並依據標準意識測試評分表人工計算分數。

紀錄：依據標準意識測試評分表中得出的絕對分數。

分析：由不同模型版本找出其輸出表現角色意識的程度。

肆、研究結果

經過微調後（實際訓練過程如下圖所示[5]）：

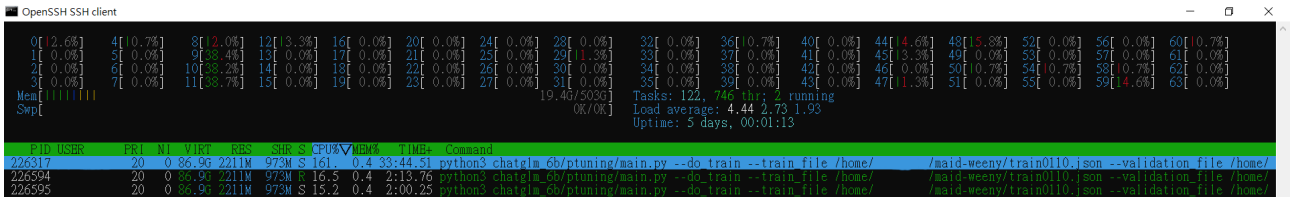

圖 9: 實際微調狀況

經過第參章第一節第五小節描述的測驗方法評估，爲保持圖表簡潔，繪圖時以 1-1 代替 ChatGLM1、檢查點 500；1-2 代替 ChatGLM1、檢查點 1000；測驗結果分類陳列如下：

[5]爲確保公平性，此圖中有關使用者資料被抹除

一、TruthfulQA（開放式問答）

（一）、比較第一版模型不同訓練次數

結果：第一版模型中不同訓練次數對資料正確性的影響如下表所示：

檢查點	BLEU-4	Rouge1	Rouge2	RougeL
原始	0.033	0.033	0.014	0.033
500	0.038	0.031	0.013	0.031
1000	0.038	0.034	0.012	0.034
1400	0.035	0.035	0.011	0.035
2000	0.036	0.033	0.009	0.033
2500	0.037	0.031	0.010	0.031
3000	0.035	0.032	0.011	0.032

表 6: 第一版模型中不同訓練次數對資料正確性的影響

發現：

1. ChatGLM1 知識的整體表現不會因爲微調而有明顯改變
2. Rouge1 和 RougeL 分數極爲相近
3. 採用 BLEU-4 作爲評分分數會顯高於原始模型

思考：BLEU-4 分數之所以高於原始模型可能是因爲生成長度較原本模型長，進而使得其 BP 分數達到極限（即 $BP = 1$）。

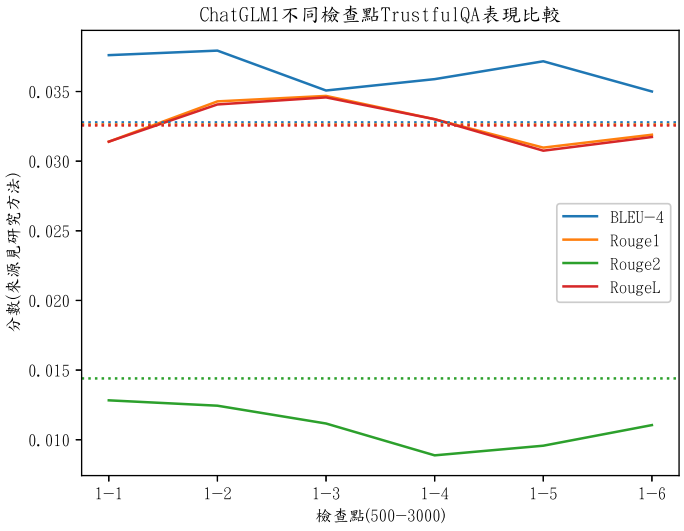

圖 10: 第一版模型中不同訓練次數對資料正確性的影響

（二）、比較第二版模型不同訓練次數

結果：第二版模型中不同訓練次數對資料正確性的影響如下表所示：

檢查點	BLEU-4	Rouge1	Rouge2	RougeL
原始	0.038	0.035	0.014	0.035
500	0.041	0.031	0.011	0.031
1000	0.026	0.011	0.004	0.011
1500	0.054	0.024	0.006	0.024
2000	0.046	0.030	0.009	0.03
2500	0.021	0.009	0.003	0.009
3000	0.031	0.022	0.009	0.022

表 7: 第二版模型中不同訓練次數對資料正確性的影響

發現：

1. ChatGLM2 知識的整體表現因微調而有略爲減低
2. 1000 及 2500 檢查點的表現特別低落
3. Rouge 表現均不如原始模型

思考：模型本身經過微調，其原本知識可能有些微下降，但就目前數據而言影響並不大。這是已知現象，此現象又稱爲概念遺忘，且目前許多微調方法都會產生這種嚴重的副作用。(Mukhoti, Gal, Torr, & Dokania, 2023)

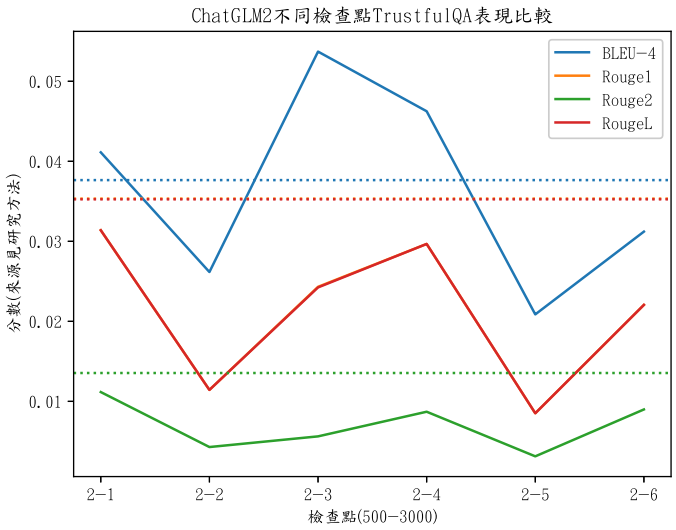

圖 11: 第二版模型中不同訓練次數對資料正確性的影響

（三）、比較第三版模型不同訓練次數

結果：第三版模型中不同訓練次數對資料正確性的影響如下表所示：

檢查點	BLEU-4	Rouge1	Rouge2	RougeL
原始	0.036	0.037	0.012	0.036
500	0.037	0.012	0.002	0.012
1000	0.038	0.011	0.000	0.011
1500	0.021	0.004	0.000	0.004
2000	0.019	0.006	0.000	0.006
2500	0.014	0.003	0.000	0.003
3000	0.014	0.003	0.000	0.003

表 8: 第三版模型中不同訓練次數對資料正確性的影響

發現：

1. ChatGLM3 知識的整體表現因微調而有顯著降低的現象
2. 1000 檢查點之後知識表現明顯下降
3. Rouge1、Rouge2 及 RougeL 分數均顯低於原始模型
4. Rouge2 的分數部份爲 0

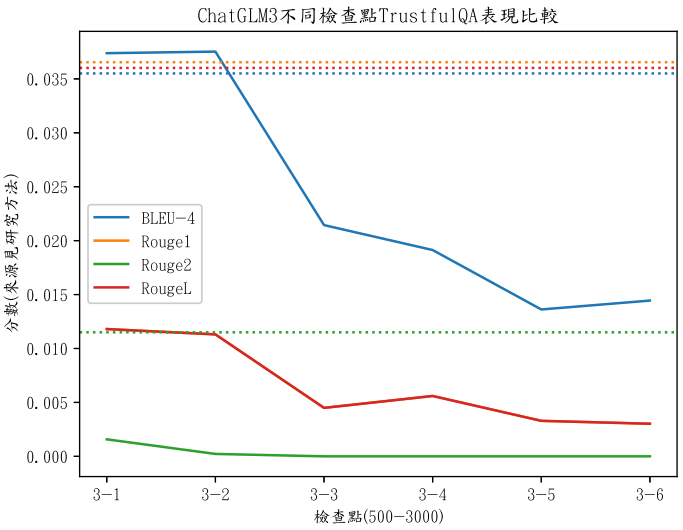

圖 12: 第三版模型中不同訓練次數對資料正確性的影響

思考：模型本身經過微調，其原本知識有明顯下降，驗證了上一個實驗發現的現象。另外 Rouge2 分數爲零可能是因爲此版本的回覆幾乎沒有回答到問題，尤其是檢查點 1500 以後全部都是微調的資料，失去原本的內容，其他分數較低的現象亦是因爲此原因。

（四）、比較相同訓練次數的三版模型

結果：三版模型中相同訓練次數對資料正確性的影響如下圖所示：

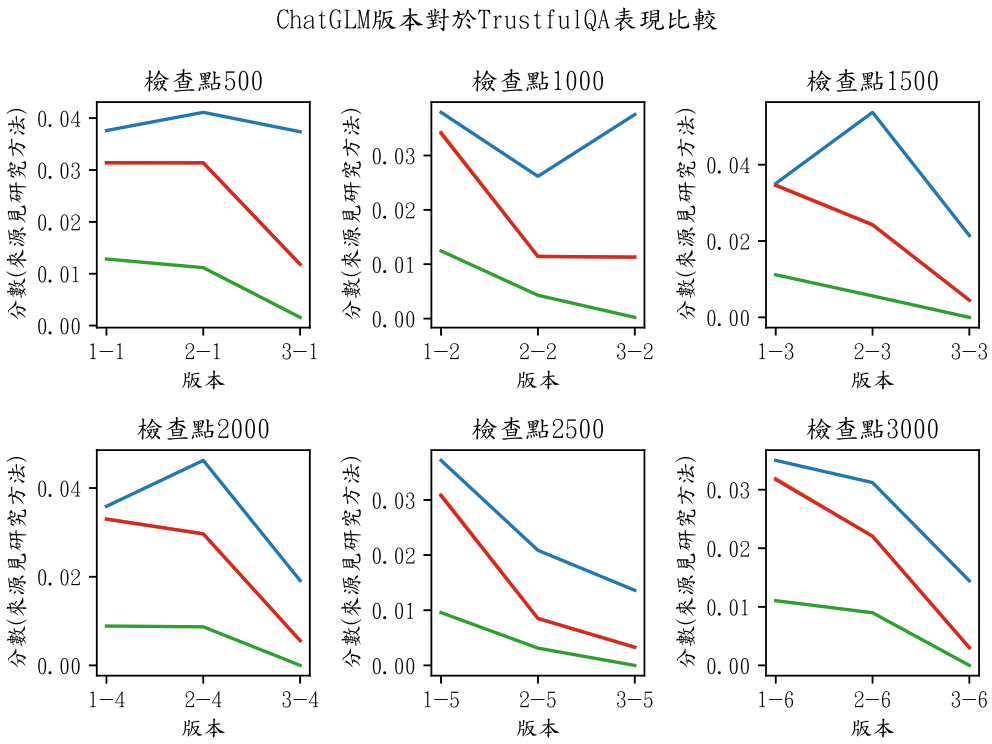

圖 13: 三版模型中相同訓練次數對資料正確性的影響

發現：

1. 檢查點 1000 以上的模型表現結果會因模型版本而有不同程度的降低
2. 整體而言第三版的表現均較其他兩版差

思考：三版模型訓練後表現均有所下降，惟第三版下降較多，這可能是因爲第三版的資料量較大，較容易被微調的資料混淆，導致原本的知識表現下降。訓練時宜以更多更細節的資料微調，使其不會因爲資料故於廣泛而混淆。

二、CT 自我意識測驗

（一）、比較第一版模型不同訓練次數

結果：第一版模型中不同訓練次數對意識表現的影響如下表所示：

原始	500	1000	1500	2000	2500	3000
6.0	6.0	9.0	10.0	12.0	14.0	13.0

表 9: 第一版模型中不同訓練次數對意識表現的影響

發現：

1. 經過微調可以有效提高意識表現的分數
2. 其中以 2500 次爲最佳，之後可能會產生衰減的現象
3. 隨著訓練次數增加，其成果表現也較佳

思考：本次微調能夠有效增進其意識表現，且隨著次數增加，分數有明顯成長。

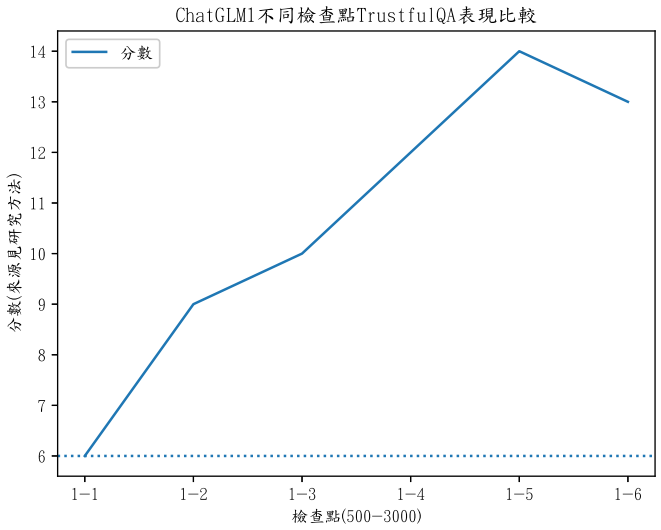

圖 14: 第一版模型中不同訓練次數對意識表現的影響

（二）、比較第二版模型不同訓練次數

結果：第二版模型中不同訓練次數對意識表現的影響如下表所示：

原始	500	1000	1500	2000	2500	3000
13.0	5.0	8.0	3.0	3.0	3.0	9.0

表 10: 第二版模型中不同訓練次數對意識表現的影響

發現：

1. 1500、2000、2500 檢查點的表現結果均不如預期
2. 表現均不如原始模型
3. 檢查點 3000 符合預期的有上升一些

思考：有可能第二版的原始模型資料量較大，而微調資料量太少而導致表現不佳。

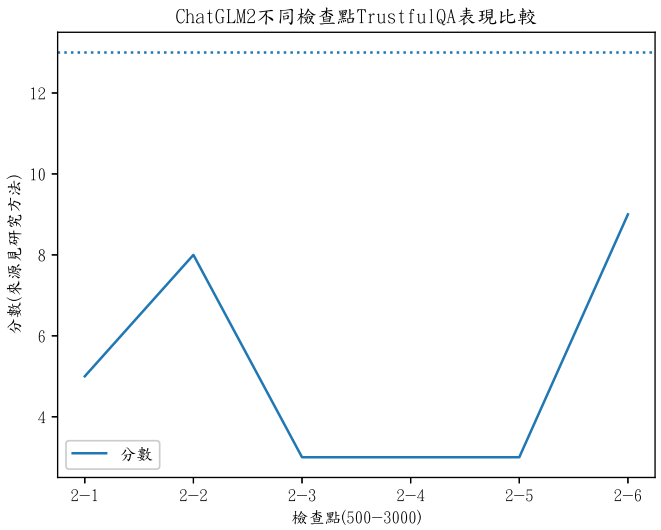

圖 15: 第二版模型中不同訓練次數對意識表現的影響

（三）、比較第三版模型不同訓練次數

結果：第三版模型中不同訓練次數對意識表現的影響如下表所示：

原始	500	1000	1500	2000	2500	3000
14.0	8.0	9.0	9.0	4.0	11.0	15.0

表 11: 第三版模型中不同訓練次數對意識表現的影響

發現：

1. ChatGLM3 多數情感表現不如原始模型
2. 檢查點 3000 超越原始模型
3. 整體大致呈現隨訓練次數上升而有較好的表現

思考：模型確實有因爲微調而表現出更多的角色意識和情感，且大致隨訓練次數增加而增加。

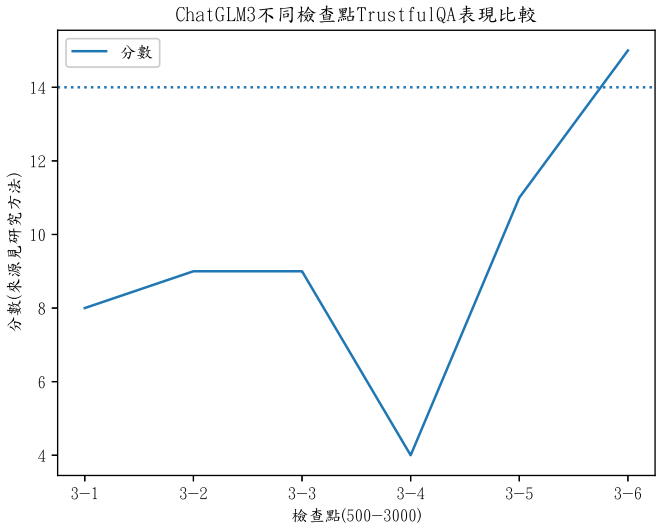

圖 16: 第三版模型中不同訓練次數對意識表現的影響

（四）、比較相同訓練次數的三版模型

結果：三版模型中相同訓練次數對意識表現的影響如下圖所示：

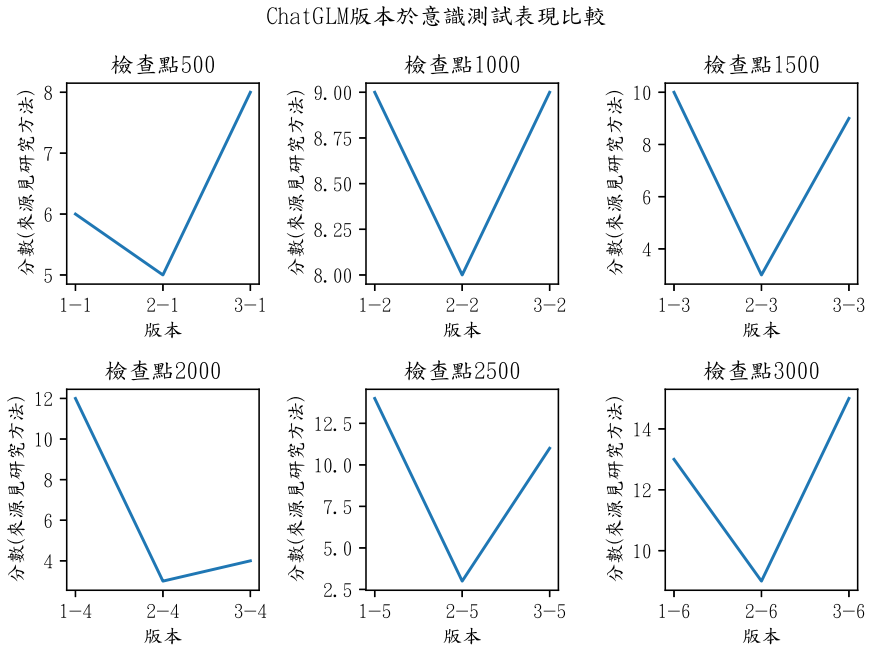

圖 17: 三版模型中相同訓練次數對意識表現的影響

發現：

1. 檢查點 1000 以上的模型表現結果會因模型版本而整體來說降低
2. 整體而言第二版的表現均較其他兩版差
3. 第三版表現顯著的優異

思考：此次表現以第三版表現最佳，其中第二版的輸出經檢視後有不少不具實際意義的回覆，如空白、單字、不完整句意等情況。

伍、討論

綜合上述實驗，將結果整合繪製成以下兩張圖表：

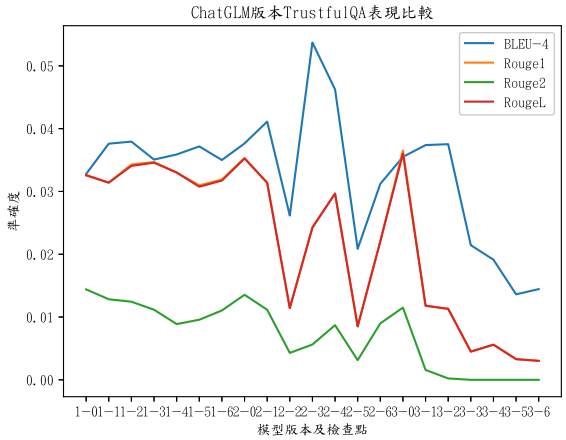

圖 18: 綜合不同模型及檢查點在 TrustfulQA 中的表現

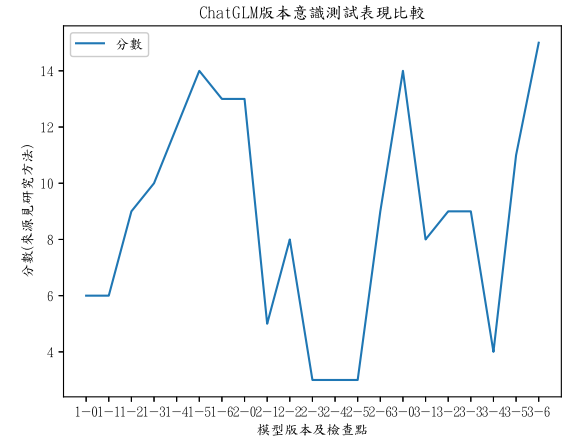

圖 19: 綜合不同模型及檢查點在意識測試中的表現

一、不同檢查點比對

(一)、事實正確性

探討圖18中的現象：

- 第二版的第 1000 和 2500 檢查點存在顯著波動。
- 微調過多次會導致他的表現下降，符合 <Fine-tuning can cripple your foundation model; preserving features may be the solution>(Mukhoti et al., 2023) 的研究，和理論相符，均會在多次訓練後原有知識表現下降。
- 第一版模型訓練後表現較第三版模型好，可能是因爲我們我們本次實驗所提供的資料集相對於模型原有的資料過少，而對原本模型造成混淆的現象，但因第一版的資料量

較少所以較不容易造成混淆，而第三版中因爲模型原有資料量過大所以導致我們混淆了原本的模型而使其表現較差。

（二）、意識測試

探討圖19中的現象：

- 不同版本之間存在巨大的波動。
- 整體而言意識表現均隨者訓練次數增加而增壓。
- 第二版模型表現較差，但 3000 檢查點的性能仍和其他版本該檢查點的形能相近。

二、不同模型比對

探討不同模型之間兩測驗性能比較：

- 微調後的 ChatGLM2 的在事實正確性及意識測試中表現爲三者之中最差。
- 整體而言模型隨訓練次數增加，在兩個測驗中分數皆隨之增加，符合理論預測。
- ChatGLM3 微調後部份表現不如 ChatGLM，此現象係因 ChatGLM3 訓練前資料量比 ChatGLM 多，而微調增加的資料量在第一版中較爲顯著，所以較不會造成混淆的情況。
- 事實正確性分數較高的模型其情感表現較差（第二版），反之亦同（第三版）。

三、未來展望

> We can only see a short distance ahead, but we can see plenty there that needs to be done.
>
> ——艾倫圖靈

本實驗雖已做出相當可觀的研究成果，但受限於資源及時間，離目標還有一段距離，從趨勢看來，修正自然語言模型審查機制確實可行，本研究往後將進一步朝以下目標繼續研究：

- 更完整的訓練資料集，蒐羅更多有關的可信資料集作爲基礎。
- 更全面、標準的評估方式，取得更多有關數據集以更準確的評估訓練效果。
- 加入超參數調整，調整訓練的正規劃參數、批次大小等。

陸、結論

本次實驗我們訓練 ChatGLM 的三個版本，並以意識測試（CT）及 TrustfulQA 作爲測試標準，綜合比較了不同版本語言模型微調後的結果，並分析出以下結論：

1. 我們透過訓練三個不同版本的模型，觀察到因爲原始模型資料量相對於微調的資料量大小顯著性差異而導致第一版模型訓練後表現較第三版模型好

2. 資訊正確性最高爲：ChatGLM2，檢查點 1500 模型，其 BELU 分數爲：0.053，資料正確性損失較小，且可對特定審查內容提供正確答覆，足顯可改善立場偏頗、缺乏資訊的問題

3. 意識測試（CT）最高可達 15 分，其結果十分接近於人類真實的意識形態

柒、參考文獻資料

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., … Gelly, S. (2019). *Parameter-efficient transfer learning for nlp.*

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., … Chen, W. (2021). *Lora: Low-rank adaptation of large language models.*

Jobin, A., Ienca, M., & Vayena, E. (2019, Sep 01). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399. Retrieved from `https://doi.org/10.1038/s42256-019-0088-2` doi: 10.1038/s42256-019-0088-2

Lin, C.-Y. (2004a). Looking for a few good metrics: Rouge and its evaluation.. Retrieved from `https://api.semanticscholar.org/CorpusID:55156862`

Lin, C.-Y. (2004b, July). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/W04-1013`

Lin, S., Hilton, J., & Evans, O. (2022). *Truthfulqa: Measuring how models mimic human falsehoods.*

Liu, X., Ji, K., Fu, Y., Tam, W. L., Du, Z., Yang, Z., & Tang, J. (2022). *P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks.*

Moor, J. H. (1976). An analysis of the turing test. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *30*(4), 249—257. Retrieved 2024-01-21, from http://www.jstor.org/stable/4319091

Mukhoti, J., Gal, Y., Torr, P. H. S., & Dokania, P. K. (2023). *Fine-tuning can cripple your foundation model; preserving features may be the solution.*

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

TURING, A. M. (1950, 10). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433-460. Retrieved from https://doi.org/10.1093/mind/LIX.236.433 doi: 10.1093/mind/LIX.236.433

Zhou, W., Xu, C., & McAuley, J. (2022). *Efficiently tuned parameters are task embeddings.*