國立臺北科技大學

資訊工程系碩士班

碩士學位論文

基於輕量化微調大型語言模型的兩階段簡繁轉換框架

Two-Stage Simplified to Traditional Chinese Conversion Framework Based on Parameter-Efficient Fine-Tuning of Large Language Models

研究生：陳禹彤

指導教授：王正豪　博士

中華民國一百一十三年七月

國立臺北科技大學
研究所碩士學位論文口試委員會審定書

本校 資訊工程系 研究所 陳禹彤 君

所提論文，經本委員會審定通過，合於碩士資格，特此證明。

學位考試委員會

委　員：楊凱翔

劉建良

王正豪

指導教授：王正豪

所　長：劉建宏

中 華 民 國　一百一十三　年　七　月　一　日

摘要

論文名稱：基於輕量化微調大型語言模型的兩階段簡繁轉換框架

頁數：39

校所別：國立臺北科技大學　資訊工程系碩士班

畢業時間：一百一十二學年度 第二學期

學位：碩士

研究生：陳禹彤

指導教授：王正豪　博士

關鍵詞：大型語言模型、輕量化微調、附加器、自然語言處理、簡繁轉換

本研究提出了一個兩階段的簡體轉繁體中文的字對字轉換架構，包含了輕量化微調大型語言模型以及非對稱映射機制，旨在提高資源受限的單機環境下的簡繁轉換正確性。在現有的方法中，有以統計式機器學習模型輔以斷詞的方式，也有先用各種方式抽取文字特徵後再以對數線性模型預測的方式，統計式機器學習模型需要極大量的語料且由於要從頭訓練所以訓練時間較久，而對數線性模型的做法則因為不是傳統自然語言處理會使用的方法所以較不直覺。因此本篇論文透過量化低秩適應附加器技術分別對簡體中文和繁體中文的大型語言模型進行指令微調。但是我們發現直接使用大型語言模型輸出的結果可能存在輸入與輸出不一致的「幻覺」問題。因此我們提出了一種能夠有效利用簡繁字之間的關係來做非對稱映射的機制，以便自動糾正這種幻覺現象。

我們利用公開的《教育部重編國語辭典修訂本》資料，依照 Alpaca 格式進行資料清洗，建立了擁有 5 個子任務及 36 萬筆指令的繁體中文指令資料集 MOE-RMCD。實

驗結果表明，簡繁轉換確實有其專業性，通用的語言模型難以泛化。本研究提出的非對稱映射機制對於未微調的大型語言模型進行零樣本的簡繁轉換任務具有非常顯著的幫助，最多能夠直接提升約 80% 的簡繁字對轉換的 BLEU 與 GLEU 分數。此外，無論是未微調的或是微調後的大型語言模型，在自然語言的理解上都優於抽取文字特徵後再以對數線性模型預測的方式約 10% 至 11%。這顯示了我們所提框架的有效性。

ABSTRACT

Title: Two-Stage Simplified to Traditional Chinese Conversion Framework Based on Parameter-Efficient Fine-Tuning of Large Language Models
Pages: 39
School: National Taipei University of Technology
Department: Computer Science and Information Engineering
Time: July, 2024
Degree: Master
Researcher: Yu-Tung Chen
Advisor: Jenq-Haur Wang Ph.D.

Keywords: Large Language Model, Parameter-Efficient Fine-Tuning, Adapter, Natural Language Processing, Simplified-Traditional character conversion

This study proposes a two-stage character-level conversion framework from Simplified Chinese to Traditional Chinese, incorporating lightweight fine-tuning of large language models and an asymmetric mapping mechanism. The aim is to enhance the conversion accuracy in resource-constrained standalone environments. Existing methods either rely on statistical machine learning models combined with word segmentation or extract textual features using various techniques before applying logistic regression models. Statistical machine learning models require a substantial amount of corpora and lengthy training times due to training from scratch, whereas logistic regression models are less intuitive as they are not traditional methods in natural language processing.

In this paper, we employ quantized low-rank adaptation （LoRA） techniques to fine-tune large language models for both Simplified and Traditional Chinese. However, we observed that the direct output from large language models might suffer from the "illusion" problem, where the input and output are inconsistent. To address this issue, we propose an asymmetric mapping mechanism that effectively leverages the relationship between Simplified and Traditional characters to automatically correct these inconsistencies.

We utilized the publicly available data from the "Revised Mandarin Chinese Dictionary" by the Ministry of Education, cleaned the data following the Alpaca format, and established a Traditional Chinese instruction dataset MOE-RMCD with five sub-tasks and 360,000 instructions. Experimental results indicate that Simplified-to-Traditional Chinese conversion indeed requires specialized handling, and general-purpose language models struggle to generalize effectively. Moreover, the proposed asymmetric mapping mechanism significantly aids zero-shot Simplified-to-Traditional Chinese conversion tasks for un-tuned large language models, improving the BLEU and GLEU scores of character conversion by up to 80%.

Additionally, whether using un-tuned or fine-tuned large language models, our approach outperforms the method of extracting textual features followed by logistic regression prediction in natural language understanding by approximately 10% to 11%.

誌謝

「充實」—— 是我對這兩年下的註腳。比起無聊而封閉的大學生活，我在研究所階段可以說是奮力生存，經歷了過往根本無法想像的有趣體驗，也幸運的登上了更高的地方並享受到了若是只甘願停留在山腰上的話絕對不可能看到的風景。

這是我人生一個相當重要的階段，我很高興兩年前的我有鼓起勇氣申請研究所，才能認識這麼多優秀且無私的朋友。

我非常感謝我的家人和厚勤，總是如此關心著我並給予我極大的支持。同時也要感謝我的指導教授給予我接近百分百的研究自由，讓我可以在這段時間盡情的學習、研究、實習以及做自己喜歡的事情，最後也要感謝實驗室同學們—于傑、子毅、致杰、永合以及念芹的互相扶持，還有優秀的學弟學妹扛下了很多很多的計畫。即使這兩年每天都像是日月輪轉般的匆忙，我依然甘之如飴。

希望我的未來也能夠像此刻的想像般，能夠百分百專注在自己喜歡的事上，並在我熱愛的所有領域中都發光發熱，燦爛無比。

陳禹彤 謹誌

中華民國一百一十三年七月

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究背景與動機

隨著大型語言模型（Large Language Model, LLM）的快速發展，其在自然語言處理（NLP）的各個領域，包括文本生成、情感分析和機器翻譯等方面，都展現出了強大的潛力（例如 OpenAI 的 ChatGPT[2] ）。然而，相對於簡體中文的百花齊放，主要語言為繁體中文的大型語言模型的數量仍然有待提高。其根本的原因就是因為使用簡體中文的人口相較於使用繁體中文的人口基數差距實在過於懸殊，尤其是對台灣而言，簡繁中文之間的轉換成為了重要的需求。因此本研究將探討如何通過微調這些大型語言模型，以提高簡繁字體轉換的正確性。

例如由中研院（2023）[1] 發布的明清歷史研究模型 CKIP-Llama-2-7b 因為使用了中國科技公司的微調模型做為預訓練模型（Pretrained Model）便導致其政治立場不如預期。但若能開發出一種更加精準的簡繁轉換方法，以將簡體中文的龐大語料轉換為繁體中文，便能一定程度上的解決台灣本土優質繁體中文語料過於稀缺的問題，因此本研究希望能以大型語言模型為底解決簡轉繁的問題。

1.2 研究目的

本研究將探討如何利用大型語言模型理解上下文關係的優勢，以改進簡繁字體轉換的正確性。我們希望在使用優質指令資料集（Instruction Dataset）語料進行輕量化微調後，使其能夠更準確地理解傳統規則為主的方法所無法涵蓋的非對稱簡繁字 [11] 轉換，從而提高簡繁字體轉換的準確性。

本研究將會嘗試解答以下問題：

1. 微調後的模型在簡繁轉換的正確性上是否有所提高？
2. 如何克服大型語言模型令人詬病的幻覺（hallucination）[3] 現象並確保模型的最終輸出（output）與輸入（input）字面上一致？

1.3 研究貢獻

這是國內首篇聚焦於應用大型語言模型進行簡繁轉換的論文。這項研究為國內在該領域的學術研究提供了新的視角，為後續研究提供了基礎和參考。

其次，我們也引入了一項精心設計的繁體中文指令資料集 — MOE-RMCD，它擁有五種子任務 —簡繁轉換、詞語解釋、單句釋義、近似詞與反義詞。這一資料集的加入，為模型對於繁體中文的理解、簡繁轉換以及字詞釋義的精準度都提供了重要的支援。

本研究的創新之處在於充分的利用了簡繁轉換的特性，從一定程度上解決了在字對字的轉換上尤其致命的幻覺問題，這不僅擴展了傳統機器翻譯的能力，而且還結合了大型語言模型在處理語言細節和上下文連貫性方面的優勢，因此可以視作為根據語境的碼對與字對轉換[19]。

此外根據實驗結果，本研究也證實了只需要再附加上其一，使用消費級顯卡即可運行的大型語言模型的輔助；其二，特定的機制，就可以在極度要求翻譯正確性的場景下成功提升開源簡繁轉換軟體的準確度，這對於低成本與低資源的研究上無疑是相當受用的。

1.4 名詞釋義

漢字，作為東亞文化的重要載體，其書寫形式的變化引起了各界的廣泛關注。1964 年，中華人民共和國對傳統漢字進行了簡化，形成了簡化字（Simplified Chinese Characters）[4]，從而產生了兩種不同的書寫系統。而隨著時代變遷，簡化字被民間稱為「簡體字」，傳統漢字則被視做筆畫較多的「繁體字」。

事實上，在台灣對於應稱「繁體字」或是「正體字」也有一定爭論 [5]，然而，即使是參閱我國政府的規章文件也無對此定一規範稱呼，例如在現行法律上曾出現過正體中文（Traditional Chinese language）[6]、正體字（Traditional Chinese Characters）[7] 與繁體中文（Traditional Chinese）[8] 的字樣，而教育部則通常稱之標準字 [9][10]。因此，本研究將採用這兩種漢字的通俗稱謂作為以下論述的基礎—「簡體」與「繁體」。

第二章 相關研究

本章會介紹過去關於簡體中文轉繁體中文的研究，以及說明大型語言模型的演進與過往語言模型的不同之處。儘管現有的轉換工具提供了一定的便利，但在準確性和效率方面仍存在不少挑戰。

2.1 簡體與繁體

Halpern 等人（1999）[19] 將簡繁轉換的層級分為了碼對的（Code）、字對的（Orthographic）、詞對的（Lexemic）以及語境的（Contextual）。例如將「计算机」轉換為「計算機」是碼對或字對的，若是轉換為台灣常用的「電腦」就是詞對的。在他們的語料庫中，有 21% 的字詞是非對稱簡繁字 [11]，這驗證了如果單只是將簡繁字之間做一個對照表是無法獲得令人滿意的結果的。

Li 等人（2010）[11] 首先使用了「非對稱簡繁字」一詞來形容簡體字與繁體字的一對多關係，例如「发」這個簡體字可能出現在「发生」與「头发」等詞語中，「發生」和「頭髮」所對應的「發」跟「髮」壓根是不相關的詞，所以「发」就是一個非對稱簡繁字。

教育部則在 2011 年發布了《標準字與簡化字對照手冊》 [10]，這本手冊的標準字以教育部「常用國字標準字體表」為範圍，共收錄 4808 個常用的標準字 [9][10] 以及與其對應的簡化字 [4]。

2.1.1 基於統計式語言模型的簡轉繁架構

Li 等人（2010）[11] 基於傳統簡繁轉換系統架構，利用統計式語言模型（Statistical Language Model）計算 unigram 和 bigram 的機率值，根據機率值決定如何翻譯。此外，他們還利用斷詞來輔助字與字之間的轉換，最後在非對稱簡繁字的轉換中得到 95.77% 的準確度（Accuracy）。

Chen 等人（2011）[12] 建立在 Li 等人的方法上，使用了整合多種統計特徵（feature）的對數線性模型（Log-Linear Model），這些特徵包含了語言模型本身以及語義的一致性，並且是以詞為單位，透過跨語言（Cross-Language）的語意空間估計得出的，最後獲得了 97.03% 的準確度。

Shi 等人（2013）[13] 受 Chen 等人 [12] 的對數線性模型啟發，透過文本分類、更多的資料以及更好的翻譯模型進一步改善了簡體轉繁體的準確度。

Xu 等人（2017）[15] 提到字與字之間非對稱關係的現象不僅只有在簡轉繁中存在，繁轉簡中也有但程度小的多。在對給定簡體句子進行繁體的字符級別轉換時，既包括用對應字符替換具有一對一對應關係的字符，也涉及了消除具有一對多映射的字符的部分，他們最後使用 KenLM [20] 為基底模型，開發出了基於統計的簡繁轉換暨校對的網頁介面。

2.2 大型語言模型

2.2.1 類神經網路與大型語言模型的開端

Vaswani 等人（2017）[21] 提出了基於自注意力機制（Self-Attention Mechanism）的 Transformer 架構，在改善既有注意力機制的同時還在模型中增加了多頭注意力（Multi-head Attention）、殘差連接（Residual Connection）與層歸一化（Layer Normalization）的機制，一舉解決了遞迴類神經網路（RNN）無法平行運算以及在文本過長的情況下表現不佳的問題，因此這個架構也對往後的各個深度學習（Deep Learning）領域都造成了深遠的影響。

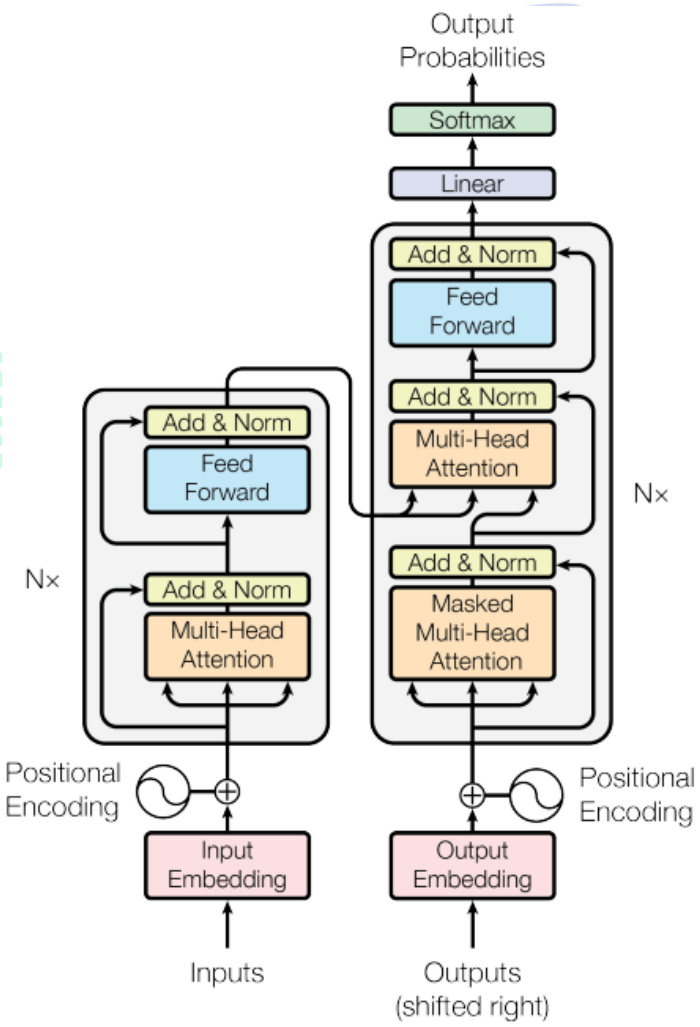

圖 2.1 Transformer 架構圖 [21]

如圖 2.1 所示。Transformer [21] 是一種 Seq2Seq 的模型，左邊是 Encoder 右邊是 Decoder，將 inputs 轉為 embedding 後會先經過位置編碼（Positional Encoding），由於 Transformer [21] 的自注意力機制不考慮單詞的順序，因此這一步驟會在 embeddings 中加入位置資訊，計算公式如下：

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

pos 是位置，i 是維度（Dimension），d_{model} 則是模型的 embedding 維度，最後他們選擇了正弦（sin）版本因為它可以推論的比原序列長度還要更長。

在自注意力機制中，對於給定的輸入，每個單詞都會計算出與序列中其他所有單詞的關聯權重並將這些做處理後輸出加權後的表示（Representation），這些多維向量代表著序列中所有單詞的資訊。

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Q、K 和 V 分別代表查詢（Query）、鍵（Key）和值（Value）的矩陣，d_k 則是鍵向量的維度。而 Transformer [21] 使用的具體來說是縮放點積注意力（Scaled Dot-Product Attention）。

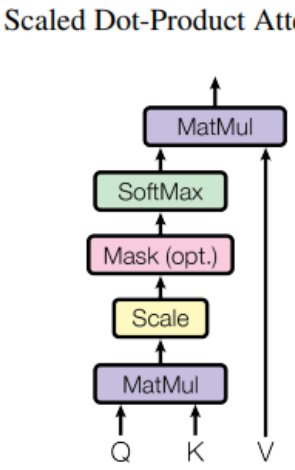

圖 2.2 縮放點積注意力機制圖 [21]

我們可以根據公式與圖 2.2 逐一拆解詳細步驟如下：

1. 計算 Query、Key 與 Value：對於序列中的每個單詞都會使用不同的權重矩陣將其轉換為 q、k 和 v 三種表示，Query 代表要注意的單詞本身，Key 是被注意的對象，而 Value 則是實際的文字內容。
2. 計算關聯度：為了增加運算速度，Q、K 和 V 就是 q、k 和 v 的矩陣版，這階段會透過計算 Query 與所有 Key 的點積（Dot Product）來衡量各單詞之間的關聯度。
3. 正規化（Normalization）：再將這些關聯度使用 Softmax 函數來轉成和為 1 的機率值，這些機率值就代表各單詞對當前單詞的重要程度。
4. 產生加權後的 Value：用上一步的機率值對所有 Value 進行加權求和，從而產生當前單詞的輸出表示。
5. 組合序列：對序列中每個單詞重複進行上述過程，最後再將所有的單詞組合起來形成此序列的表示。

多頭注意力則是 Transformer [21] 的核心，它可以讓模型在處理序列資料時同時關注序列中的多個位置，這種機制除了可以讓計算平行化之外也解決了自注意力機制可能只看到序列中某些部分的問題。

從圖 2.3 可以看到，多頭中的「頭」就是被分割成較小維度的嵌入向量，這些頭會被平行的處理，也就是進行自注意力權重的計算，最後會將這些頭的輸出向量串聯起來做線性變換（Linear Transform）後生成最後的輸出。

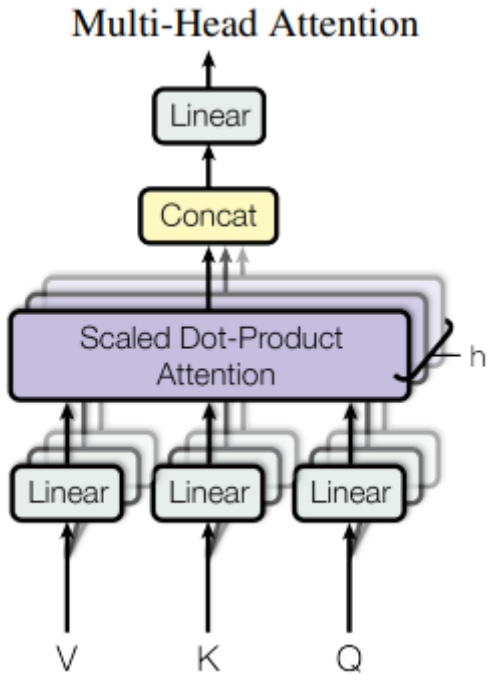

圖 2.3 多頭注意力機制圖 [21]

現代流行的大型語言模型幾乎都使用到了 Transformer 的架構。Agarwal 等人（2018） [3] 探討了類神經機器翻譯（NMT）系統中的一種特殊現象，稱為「幻覺」。這個現象不常發生，但是這些翻譯結果與原始文本完全脫節，嚴重影響使用者對系統的信任，在他們的實驗中有 48%～73% 的句子會受到幻覺的干擾。他們的分析顯示，普通的翻譯結果與有幻覺的結果在注意力矩陣（Attention Matrix）上有著根本上的差異，這可能是和解碼器（decoder）忽略了編碼器（encoder）的上下文有關。

2.2.2 預訓練語言模型的新範式

與傳統以 BERT 和 Encoder-Only 為主的語言模型截然不同，基於轉換器的生成式預訓練（Generative Pretrained Transformer, GPT）流派開始悄悄興起，這種 Decoder-Only 的語言模型捨棄了繁複的遷移學習（Transfer Learning）的過程轉而投向提示工程（Prompt Engineering）的懷抱。

Brown 等人（2020）[31] 共同訓練了擁有 1750 億參數的 GPT-3，他們的模型以及思想可以說是顛覆了傳統語言模型的範式（paradigm）。他們將模型在無監督式（Unsupervised）的預訓練過程中發展出的可在推理時快速適應或識別任務的能力稱之為上下文學習（In-Context learning, ICL）。因此 GPT-3 無須任何參數更新或微調即可透過使用者給定的提示詞（prompt）來進行任務，例如零樣本（zero-shot）、單一樣本（one-shot）或是少量文本（few-shot）。

Wei 等人（2021）首先提出了指令微調（Instruction Tuning）[25] 的概念。該研究探討了對大型語言模型進行指令微調的方法與效果，他們將 62 種資料集分類成 12 種子任務類別並使用 137B 的 LaMDA-PT 作為基底模型，實驗證實在經過指令微調後，原先的預訓練模型在各種任務上的零樣本效能都顯著提高了。

後來的 Taori 等人 （2023） 也受 Wei [25] 等人影響，發布了基於 LLama-7B [26] 模型的 Alpaca-7B [22]。他們借鑑了自我指導（Self-Instruct）的概念並使用 OpenAI 的 text-davinci-003 的模型自動生成 52K 的指令資料集（Instruction-following dataset），再進行基於指令遵循（Instruction-following）範式的微調。令人驚訝的是，這種方式相較於傳統人工標註而言，以低成本（少於 500 美金）與極小模型尺寸（7B v.s. 175B）的情況下有效的復刻了原始的 Text-davinci-003 的效能。

2.2.3 輕量化微調

漸漸的，全參數微調（Full Parameters Fine-Tuning）已經不是普通公司吃得消的技術，因此針對低資源情境也有一些研究出現，也就是輕量化微調（Parameters-Efficient Fine-Tuning），其中效果最優越的當屬附加器（Adapter），透過在預訓練模型的各層中

插入附加器，可以達到快速且高效的調整權重與偏置 (bias) 的效果，最後再融合調整好的參數回該層就可完成整個微調流程。

Hu 等人（2022）提出了低秩適應（Low-Rank Adaptation, LoRA） [24] 如圖 2.4 [24]，這是一種透過凍結 Transformer [21] 的原始權重並在各層中加入可訓練的低秩分解矩陣以達到減少訓練參數的技術。在圖 2.4 中，左邊的 Pretrain Weights 代表原始預訓練模型的權重，它是一個 d×d 維的矩陣，右邊的黃色區塊則是由兩個低秩矩陣組成，下面的 A 矩陣的維度是 d×r，而上面的 B 矩陣則是 r×d。在微調之前，A 會被初始化成隨機的小數值，而 B 矩陣為一個數值全 0 的矩陣，在微調的過程中，模型會學習並慢慢更新 A 與 B 中的數值，使得兩矩陣的乘積逐漸逼近原始的模型權重。

這種技術相較於傳統的全參數微調來說，可以極大幅的減少訓練參數至一萬倍以及大幅減少 GPU 的 VRAM 使用量至三倍。此外，在 GPT-2 [32] 與 GPT-3 [31] 的模型中，他們方法的效果甚至超越了全參數微調。

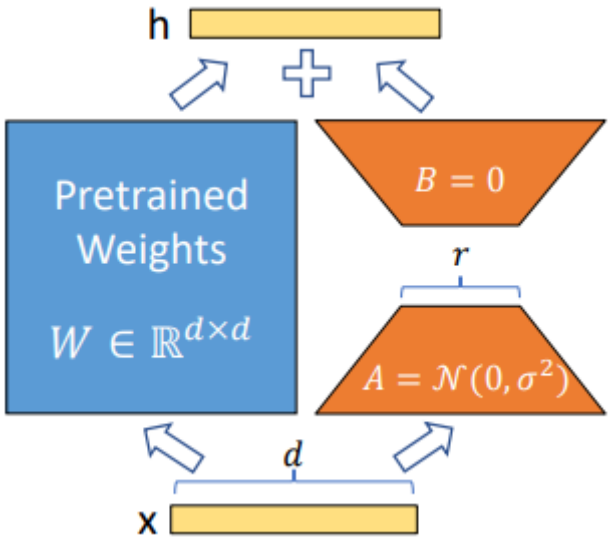

圖 2.4 低秩適應方法示意圖 [24]

Dettmers 等人（2023）基於 Hu 等人的 LoRA [24] 方法進一步提出了量化低秩適應（Quantization Low-Rank Adaptation, QLoRA）[30] — 這是一種更專注在節省 VRAM 但卻不降低效能的進階技術，為了能夠在將模型量化成 4-bit 的情況下凍結參數且使用 LoRA [24] 微調，他們做了諸如 4-bit NormalFloat（NF4）、雙重量化（Double Quantization）與分頁優化器（Paged Optimizer）的創新。此外他們也證實了對於特定任務而言，想要獲得優質模型的最重要因素是資料集的品質而非大小。

2.2.4 開源的大型語言模型

雖然 ChatGPT [2] 相當強大但是它並不是以開源的方式對外開放，因此也開始有一些公司釋出自己的研究成果。

例如由 Meta 研究員為主的 Touvron 等人（2023）介紹了 LLaMA [26]，他們除了調整超參數（hyper-parameters）之外也受先前一些模型架構的啟發，使用了對輸入進行預歸一化（Pre-normalization）、SwiGLU 激勵函數（Activation Function）以及旋轉位置嵌入（Rotary Postition Embedding）等技巧，成功在大多數的基準測試中贏過 10 倍之大的 GPT-3。最後釋出了 7B、13B 與 70B 三種不同大小的模型。

短短五個月後，Touvron 等人又發布（2023） LLaMA 2-CHAT [27]，這是 LLaMA 2 [26] 的聊天版，在預訓練基座模型時除了對資料做更好的預處理、更新使用的資料集、將訓練的 token 數量增加 40% 外也增加了上下文長度（context length）與群組查詢注意力（Grouped-query Attention, GQA） 的機制，最後再使用基於人類回饋的強化學習（RLHF）微調模型，使其輸出內容達到有幫助的（helpfulness）與安全的（safety）的雙重目標。

有鑑於主流的大型語言模型都以英文為主，也開始有人著手訓練中文的大型語言模型。

Cui 等人（2023）使用了簡體中文語料來訓練一個中文分詞器（tokenizer），再與舊有的 LLaMA [27] 分詞器合併，再用類似於 Alpaca [22] 的提示詞模板以及 LoRA [24] 方法進行指令微調 [25] 以增強 LLaMA 2 [27] 的簡體中文能力，最終釋出了 Chinese-LLaMA-2 [29] 的兩種版本— 7B 與 13B。

Lin 等人（2023）則是基於 LLaMA 2 [27] 釋出了 Taiwan-LLM [28]，這是第一個專為台灣文化設計的繁體中文大型語言模型，他們使用繼續預訓練（Continue-Pretraining, cPT）、監督式微調（Supervised Fine-Tuning, SFT） 與回饋監督式微調（Feedback Supervised Fine-Tuning, Feedback SFT） 這三種階段來達成目標，經過這些階段後的 Taiwan-LLM 與原始的 LLaMA 2-CHAT 相比，無論是在台灣文化或是情緒分析上的 benchmark 中都有著顯著的進步。

TAIDE （2023） [35] 是由國科會與其轄下的機構共同打造的「可信任生成式 AI 對話引擎」（TAIDE，Trustworthy AI Dialogue Engine）計畫的簡稱，他們希望能夠做出繁體中文專屬的可信賴人工智慧應用的基底模型，已經釋出包含根據 LLaMA 2 [27] 做微調的 TAIDE-LX-7B [36] 與 TAIDE-LX-7B-CHAT [36]，TAIDE-LX-7B-CHAT 除了指令微調 [25] 外它也有做語表擴充並針對辦公室常見應用以及台灣在地文化的知識上做加強。

第三章 研究方法

本章節將介紹此論文提出的方法與架構，這是一個兩階段式的簡繁轉換方法，可以分為簡繁轉換模型階段與非對稱映射階段。框架的架構圖如圖 3.1。

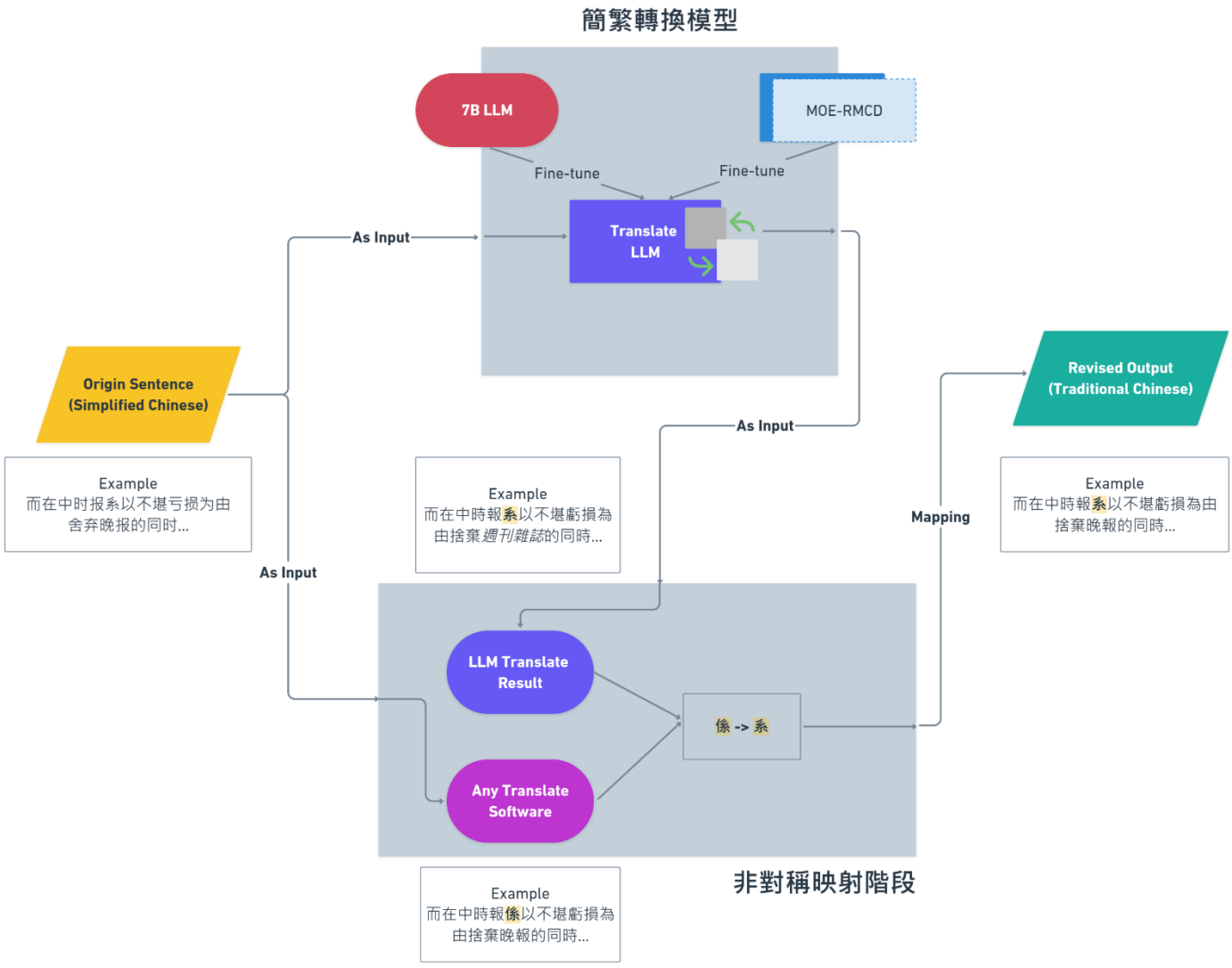

圖 3.1 兩階段簡繁轉換框架之架構圖

當我們輸入原始句子時，他會先進入簡繁轉換模型的部分，因此我們會拿到由大型語言模型輸出的參考翻譯。再來進到非對稱映射階段，除了參考句外，我們還會使用任意一個簡繁轉換軟體獲取翻譯結果做為原始翻譯，最後再使用非對稱映射機制綜

合原始翻譯與參考句的優勢輸出最終的翻譯結果。在下面的章節中，會更詳細的說明各個階段的作用以及其實作細節。

3.1 簡繁轉換模型訓練

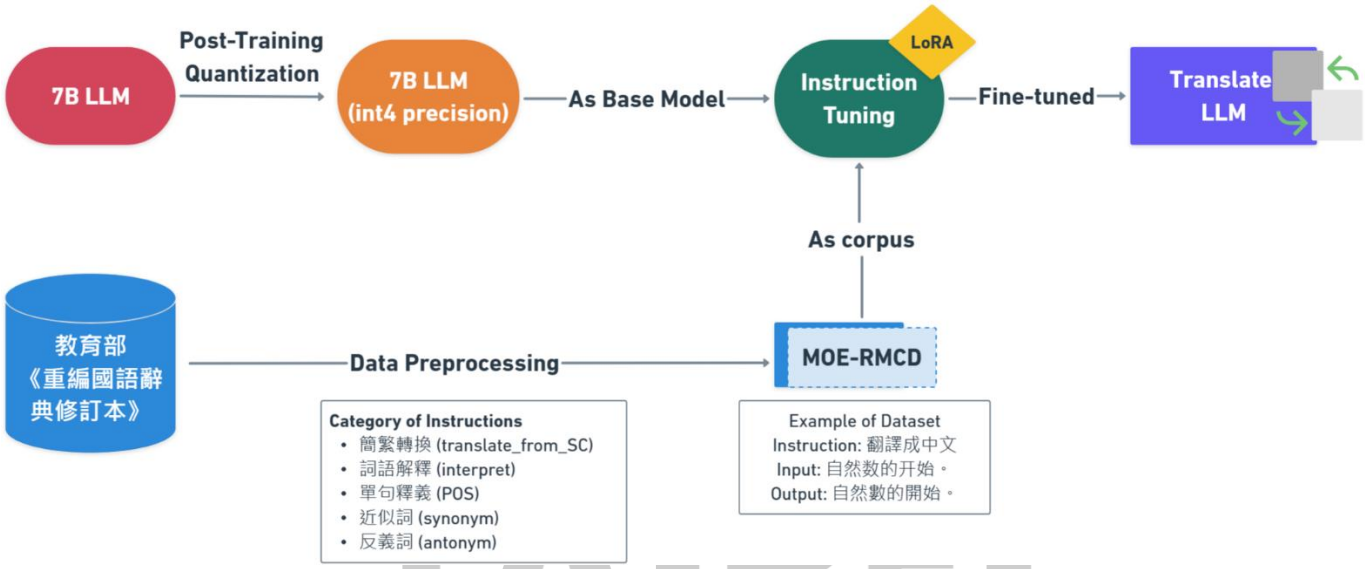

圖 3.2 簡繁轉換模型訓練架構圖

要使用簡繁轉換模型之前，我們必須先訓練出一個模型。我們使用 70 億參數的大型語言模型做為基底，並將模型量化壓縮至 int4 的準確度，由於此一步驟是在模型訓練後才量化，因此此步也被稱為訓練後量化（Post-Training Quantization, PTQ）。

得到量化後的模型後，我們再搭配我們整理好的 MOE-RMCD 指令資料集進行指令微調 [25]。由於 VRAM 的限制我們選擇使用較輕量化的方式對模型進行微調，LoRA [24] 是一種輕量化微調的技術，它可以極大幅減少訓練參數的數量，透過上述的方法，就可以得出一個專門用來做簡繁轉換的大型語言模型。

3.2 非對稱映射機制

雖然微調完的大型語言模型可以理解簡體到繁體的轉換規則，但是受到類神經機器翻譯特有的「幻覺」[3] 影響，翻譯的結果並不穩定，且可能會有多字或漏字的情況。我們以同一句輸入為例，並讓模型推論多次的結果節錄如下。

輸入：

而在中时报系以不堪亏损为由舍弃晚报的同时，另方面却持续入股中天电视台，并有意在未来收购中视，成就跨媒体集团霸业。

…

輸出二：

而在中時報系以不堪虧損為由捨棄週刊雜誌的同時，另方面卻持續入股中天電視台，並有意在未來收購中視，成就跨媒體集團霸業。

可以發現在輸出二當中「晚報」一詞被錯誤的輸出為了「週刊雜誌」，這對於簡繁轉換而言無疑是相當致命的問題，尤其在我們的命題上只有字符之間的轉換，這代表在我們的預期中，無論怎麼轉換，都不應該影響到原句與輸出的字符。

為了解決這個問題且考慮到非對稱簡繁字的特性，我們提出了非對稱映射（Asymmetric Mapping, AM）的機制。之所以命名為「非對稱映射機制」是因為這項機制只會在句子內有非對稱簡繁字 [11] 時觸發，它以傳統機器翻譯結果為原始翻譯（Raw Translation），並將模型對非對稱簡繁字的轉換作為參考（Reference）。這個機制是建立在簡繁轉換模型對於非對稱簡繁字的轉換優於 baseline 的假設下構想出來的，因此，它會使用原始翻譯做為最終翻譯結果的骨架，並將參考句中的非對稱簡繁字無條件的，直接替換掉所有原始翻譯句中的非對稱簡繁字。

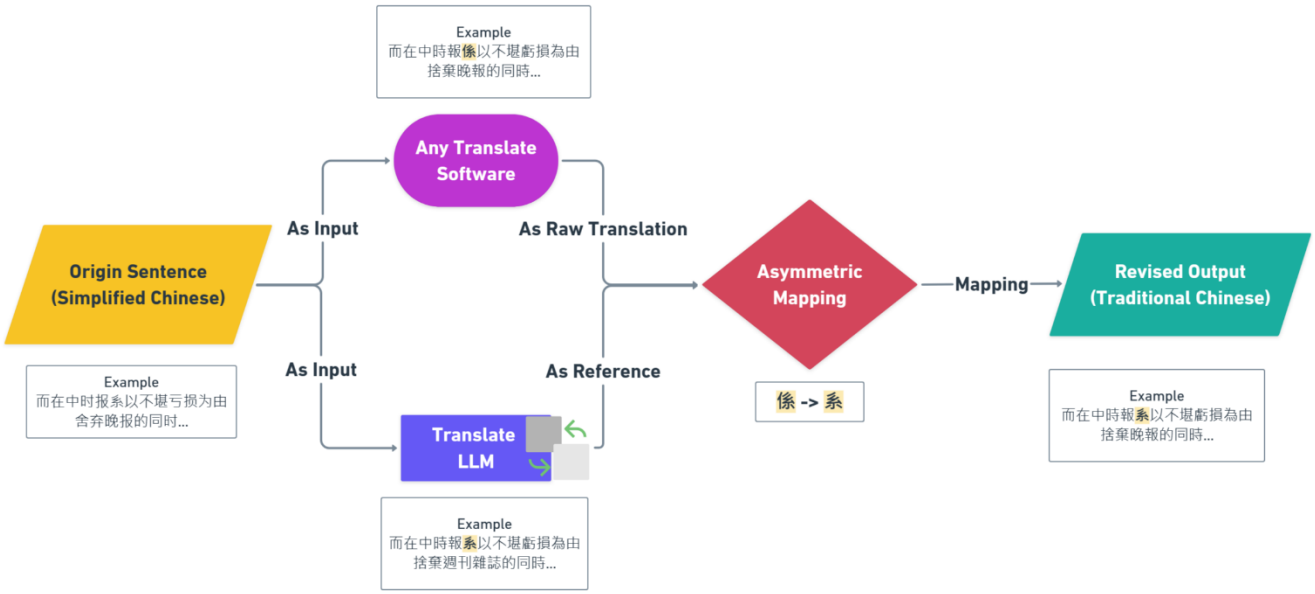

圖 3.3 非對稱映射機制運作圖

例如，我們現在有一句「而在中时报系以不堪亏损为由舍弃晚报的同时...」需要翻成繁體中文，這句會同時做為任一翻譯軟體以及大型語言模型的輸入，以 OpenCC 為例，他會將這句翻成「而在中時報係以不堪虧損為由捨棄晚報的同時...」，而 Chinese-Alpaca 則是翻為「而在中時報系以不堪虧損為由捨棄週刊雜誌的同時...」，這時候我們以翻譯軟體的輸出做為原始翻譯，並使用大型語言模型對於非對稱簡繁字的處理做為參考，直接將參考中對應的非對稱簡繁字替換到原始翻譯上。這一步可以解決模型的原始輸出可能會出現幻覺的問題，因為此舉相當於將原始句做為上下文輸入，使用大型語言模型預測，並消除該非對稱簡繁字的一對多的歧字。

如此一來，既可以保證簡繁轉換後的輸出跟輸入的字面意義完全相同，又可以運用到大型語言模型比起過去的統計式語言模型而言，更加理解上下文語義的優勢。

第四章 實驗與討論

4.1 資料集介紹

由於簡繁字體轉換的資料集取得不易，且能找到的同領域文獻年代相對久遠，導致其所使用的資料集很多都已經無法取得（例如 1998-2001 年的新聞資料庫 [11] 與中華人民共和國教育部與中國中文資訊學會在 2013 年發佈的 MOE-CIPSC [15]）或是所費不貲（例如 LDC2007T38 [12] 要價 4000 美元），因此如何在低成本且公允的情況下檢驗我們方法的可行性變成一件很難的事。

本論文引入新的資料集——「教育部重編國語辭典修訂本指令資料集」（Ministry of Education Revised Mandarin Chinese Dictionary Instruction Dataset，以下簡稱 MOE-RMCD），它是由教育部的《重編國語辭典修訂本》[18] 為底所構建的指令資料集。基於想要盡可能最大化利用原始資料潛在價值的想法，我們從中抽取出五大類任務——詞語解釋、簡繁轉換、單句釋義、近似詞與反義詞。

詞語解釋和單句釋義皆係整理原資料之「釋義」所衍伸；簡繁轉換是將無法被特別抽取的語句轉換為簡體，再將簡體做為指令輸入（instruction input），繁體做為指令輸出（instruction output）而成；近似詞與反義詞則分別來自原資料的「相似詞」和「相反詞」。下面會介紹訓練資料集的資料格式，和本論文對此資料集做的前處理過程。

而用來評估的測試集則是使用由國家教育研究院釋出的「國家教育研究院華語文語料庫」（Corpus of Contemporary Taiwanese Mandarin, COCT）[24]，此語料包括書面語、口語、華英雙語及華語中介語。我們使用的是華英雙語語料庫的中文部分，我們抽取了其中 10% 作為測試集，此測試集共有 31,091 筆。

此外，為了實作非對稱映射機制，我們必須先知道哪些字是非對稱簡繁字。所以我們將《標準字與簡化字對照手冊》[10] 中的資料轉為 json 檔，如以下範例。

```
## 對稱簡繁字範例
{
  "又": {
    "正體字": [
      "又"
    ],
    "非對稱簡繁字": false
  }
}
## 非對稱簡繁字範例
  "干": {
    "正體字": [
      "干",
      "乾",
      "幹"
    ],
    "非對稱簡繁字": true
}
```

資料中的鍵（key）為簡體字，而值（value）中的正體字則記載了所有該簡體字可以被映射到的繁體中文字，如果非對稱簡繁字一欄為 true，就代表該字是非對稱簡繁字，反之亦然。

4.1.1 訓練資料集介紹

MOE-RMCD 這個指令資料集以教育部的《重編國語辭典修訂本》[18]（以下簡稱《國語辭典》）為底，根據 Alpaca-7B [22] 的格式做了細粒度的資料清洗，它的內容包含在中日韓統一表意文字列表 （CJK Unified Ideographs） [22] 與中日韓統一表意文字擴展區 A （CJK Unified Ideographs Extension A） [22] 內的所有漢字，同時排除了過於罕見的字。處理後，此資料集共計有 365,461 筆指令對 （instructions pairs），並包含五大類任務，詳細見表 4.1。

表 4.1 MOE-RMCD 資料集任務中英對照表

詞語解釋	簡繁轉換	單句釋義	近似詞	反義詞
interpret	translate_from_SC	POS	synonym	antonym

MOE-RMCD 資料集的資訊如表 4.2 所示。此資料集共有 365,461 筆資料，其中詞語解釋任務有 171,097 筆，佔比 46.82%，而簡繁轉換的筆數有 163,692 筆，佔整個資料集的 44.79%，單句釋義、近似詞與反義詞總佔比是 8.39%。

表 4.2 MOE-RMCD 資料集統計資訊

詞語解釋	171,097
簡繁轉換	163,692
單句釋義	8,170
近似詞	13,899
反義詞	8,603
資料集總筆數	365,461

4.1.2 訓練資料集處理方法

1. 簡繁轉換（translate_from_SC）

- 任務描述: 以字符的層級而言，從繁體轉到簡體並沒有問題 [19]，所以我們將原始語言為繁體中文的國語辭典的文本轉換為簡體，再做為此 instruction 的 input。
- 來源: 國語辭典內無法被特別處理的文本。
- 資料範例 - 現代中文

Instruction: 翻譯成中文 Input: 自然数的开始。 Output: 自然數的開始。

- 資料範例 - 無法被特別處理的文本

Instruction: 翻譯成中文 Input: 1.清洁。《红楼梦》第四四回：「那市卖的胭脂都不干净，颜色也薄。」《文明小史》第五二回...。 Output: 1.清潔。《紅樓夢》第四四回：「那市賣的胭脂都不乾淨，顏色也薄。」《文明小史》第五二回...。

2. 詞語解釋（interpret）

- 任務描述: 解釋詞或給定句子求特定字的解釋。
- 來源: 根據國語辭典的「釋義」拆分而成。
- 資料範例 - 詞義

Instruction: 解釋這個詞 Input: 泵 Output: 一種用以增加液體的壓力，用來移動液體的裝置，即是對流體作功的機械。也稱為「幫浦」。

- 資料範例 - 現代中文例句

Instruction:「你是不是背著我幹了什麼壞事？」裡面的「背」是什麼意思 Input: NaN Output: 動詞，躲避、瞞著

- 資料範例 - 文言文例句

Instruction: 「《詩經・小雅・常棣》:「妻子好合，如鼓瑟琴。」裡面的「合」是什麼意思 Input: NaN Output: 形容詞，和諧、融洽

3. 單句釋義（POS）

- 任務描述: 列出給定詞語與詞性，要求解釋使用時機。
- 來源: 根據國語辭典的「釋義」拆分而成。
- 資料範例 – 單句釋義

Instruction:「旭」什麼時候可以做為這個詞性使用 Input: 名詞 Output: 剛升起的太陽。如：「朝旭」、「迎旭而舞」。

4. 近似詞（synonym）

- 任務描述: 列出給定詞語的近似詞。
- 來源: 國語辭典的「相似詞」。
- 資料範例 – 近似詞

Instruction: 舉出 1 個近似詞 Input: 大都

Output: 多數

5. 反義詞（antonym）

- 任務描述: 列出給定詞語的反義詞
- 來源: 國語辭典的「相反詞」
- 資料範例 – 反義詞

Instruction: 舉出 4 個反義詞 Input: 寬大 Output: 狹小、窄小、嚴厲、苛嚴

4.2 實驗環境與參數設定

關於實驗環境，整個實驗皆是在 Ubuntu 20.04.6 LTS 下完成，為了最大化的善用資源，我們也安裝了版號為 530.30.02 的 NVIDIA 驅動，以及相對應的 CUDA 12.1 與 CUDNN 12.1，硬體部分選擇的是 NVIDIA GeForce RTX 3080 Ti 12G 顯示卡。

使用 LoRA[24] 技術微調一個 7B 的模型需要至少 16GB 的 VRAM，所以我們必須先量化壓縮大型語言模型至 4 位元或 8 位元之後進行輕量化微調 （Parameter-Efficient Fine-Tuning, PEFT），這種過程就是 QLoRA[30] 的核心精神。

我們在將模型量化壓縮至 4 位元並搭配腦浮點數 （Brain Floating Point, BF16） 數值精度在 GeForce RTX 3080Ti 上進行訓練，相較於傳統的全參數微調，它所占用的計算資源減少了 96%，此外因應 GPU 規格限制，我們也把訓練期間能使用的最大記憶體空間設為 12GB。

將兩個低秩矩陣的秩大小（r）設為 64，參數更新速度（alpha）設為比原始參數快 16 倍以利快速收斂，學習率（learning rate）設為 0.0002，每 16 步更新一次梯度（gradient），優化器則使用經過分頁優化的 AdamW 32bits，梯度裁剪最大範數設為 0.3 以防止梯度爆炸（gradient explosion）的問題，訓練步數的話我們使用 Chinese-Alpaca-2-7B 試驗了 250 步到 10000 步之間的差異，最終選擇了 1250 步的版本。

除了微調的參數外，大型語言模型的參數也是影響翻譯品質一個很大的點。我們設定 top_k 為 50，也就是只從模型認為最可能的前 50 個選項中選擇下一個單字以增加生成文本的連貫性和可讀性。top_p 為 0.95 代表只考慮一個累積機率達到 95% 的最小的單詞，可以幫助模型在保持多樣性的同時避免生成低機率的奇怪或無關的文本。並利用 max_new_tokens 參數來控制生成的輸出不會比輸入的句子更長。

4.3 評估指標

本研究以 BLEU 與 GLEU 兩種指標進行翻譯模型的可用性評估。

BLEU（Bilingual Evaluation Understudy）是一種用於評估從一種自然語言機器翻譯到另一種語言的文本品質的指標。它通過將機器生成的文本與一個或多個人工生成的參考翻譯進行比較來計算翻譯品質。它廣泛用於自然語言處理中，以評估和改進機器翻譯模型，提供了一種量化的手段來衡量翻譯的準確性，如公式（4.1）所示。

$$\text{BLEU} = BP \cdot exp\left(\sum_{n=1}^{N} \text{w}_n \text{logp}_n\right) \qquad (4.1)$$

BP 是短句懲罰因子（Brevity Penalty），用於處理機器翻譯結果過短的情況。exp 表示指數函數，用於計算後面括號內的加權平均對數精度。N 則是代表最高考慮的 n-

gram 數量，例如 N = 4 時，這個式子就衡量 1-gram 到 4-gram 的所有 n-gram 的加權平均的和。w_n 是第 n 個 n-gram 的權重，用於確定不同長度 n-gram 的相對重要性。$logp_n$ 是第 n 個 n-gram 的對數精度，計算方式是該 n-gram 在機器翻譯結果中正確出現的次數除以在機器翻譯結果中出現的總次數的對數。

GLEU（Google-BLEU）則是 Google 基於 BLEU 指標所產生的一種變體。具體而言，GLEU 評分會在翻譯出的句子中計算所有可能的 n-gram（通常 n 的範圍是 1 到 4）的匹配數量，然後對目標句子計算 n-gram 的準確率 （Precision） 和召回率 （Recall），並取兩者的最小值。因此 GLEU 分數的範圍會在 0 與 1 之間，0 表示完全沒匹配到，1 則表示完全匹配，這種細微的改動使它可以更好地評估短語句的翻譯。

4.4 Baseline 介紹

本論文要比較的對象如下：

1. ***OpenCC*** [33]：Open Chinese Convert （OpenCC, 開放中文轉換） 於 2010 年首度發佈，是簡繁轉換領域中的相當知名的開源專案。支援字元級、詞組級轉換、漢字變體轉換以及地區之間的習慣用詞轉換，如「裏」與「裡」、「鼠標」與「滑鼠」等。
2. ***Google Translate*** [34]：由 Google 開發的免費多語言機器翻譯服務，它能提供文字、文件和網頁的即時翻譯。自 2006 年推出以來，Google 翻譯已支援超過 100 種語言。
3. ***XMUCC*** [12]：漢字簡繁文本智慧轉換系統由廈門大學智慧科學與技術系自然語言處理研究組開發。它提供了兩種介面，在網頁上可選擇做針對台灣、香港與

古籍的簡繁轉換，此外也有 Word 擴充功能以及可在 Windows 上運行的單機版程式可供下載。

4.5 實驗結果

我們首先會先比較三種 Baseline 各自的 BLEU 與 GLEU 分數，接著再使用本論文所提出的非對稱映射架構對三種大型語言模型——以繁體中文為主的 Taiwan-LLM-7B-v2.0.1-chat [28]（下稱 Taiwan-LLM）、有繁體中文詞表擴充的 TAIDE-LX-7B-Chat [35]（下稱 TAIDE）和以簡體中文為主的 Chinese-Alpaca-2-7B [29]（下稱 Chinese-Alpaca）做試驗，因為三個都是由 LLaMA-2 衍生而成。最後也會再列舉出一些在引入非對稱映射機制後的原始翻譯、參考以及最後的修正結果作為範例。

三種 Baseline 各自的 BLEU 與 GLEU 分數的結果如表 4.6 所示。OpenCC 與 XMUCC 均來自本機端的執行結果，Google Translate 則是經由 API 取得。

表 4.6 Baseline 間的分數比較結果

Baseline	OpenCC		Google Translate		XMUCC	
	BLEU	GLEU	BLEU	GLEU	BLEU	GLEU
Score	0.9391	0.9641	**0.9992**	**0.9989**	0.8420	0.9049

可以連網且有強大算力支援的 Google Translate 毫不意外地效果最好，此外得益於其開源的特性，OpenCC 比起還停留在論文初發佈時代的 XMUCC 而言效果也好上許多，即使兩者的背後原理都是基於傳統的統計式語言模型。

再來表 4.7 可以觀察到不同 Baseline 與不同大型語言模型搭配的實驗結果。

表 4.7 非對稱映射機制的實驗結果

Baseline 非對稱映射機制 （AM）	OpenCC		Google Translate		XMUCC	
	BLEU	GLEU	BLEU	GLEU	BLEU	GLEU
Baseline	0.9391	0.9641	**0.9992**	**0.9989**	0.8420	0.9049
Chinese-Alpaca （未微調）＋AM	0.9396	0.9644	0.7574	0.8099	0.8466	0.9073
Taiwan-LLM （未微調）＋AM	0.9413	0.9654	0.7643	0.8009	0.8568	0.9137
TAIDE （未微調）＋AM	0.9406	0.9650	0.9832	0.9900	0.8465	0.9074
Chinese-Alpaca （微調後）＋AM	**0.9590**	**0.9758**	0.8078	0.8717	**0.9025**	**0.9417**
Taiwan-LLM （微調後）＋AM	0.9455	0.9678	0.8220	0.8652	0.8795	0.9275
TAIDE （微調後）＋AM	0.9518	0.9716	0.9758	0.9856	0.8973	0.9386

可以發現，如果是使用單機版的翻譯軟體做為原始翻譯的情境下，無論是使用未微調，抑或是微調後的大型語言模型，在加上非對稱映射機制後，都能有效的提升 Baseline 原始翻譯的結果，也因此我們能夠推論出「未微調的大型語言模型本身理解上下文的能力就比傳統語言模型來的好」的觀點。

但是這個機制在使用 Google Translate 做為原始翻譯下反而效果會變差，因此我們能夠得知，如果想要使用非對稱映射機制，那必須保證大型語言模型的翻譯效果就比原始的 Baseline 好，否則反而會降低整體的翻譯效果。

另一個令人感到不解的點則是——在以 Google Translate 做為 Baseline 的情況下，未微調的 TAIDE 居然表現的比已微調的 TAIDE 好。為了弄清楚這個問題，我們將模型回應內容節錄如下：

原始輸入：

截至民國六十五年三月底為止，政府已經協助青年九百四十三人貸款創業，循環貸出新台幣一億零三百八十五萬四千元，所創事業全年生產額已達新台幣十五億餘元，創業的種類包括化工、機械製造、電機製造、運輸工具製造等二十八種行業。

TAIDE （未微調）:

基於上述資料，以下是模型的簡要回應：

截至民國六十五年三月底，政府已針對 943 名青年創業補助貸款共撥款 9,433 萬 8,540 元，依循循環貸款的新台幣 1 億元餘元持續挹注，該年度共達台幣 15 億元之多；創業類型包含化學、機械、電機、工具機、運輸工具等共 28 個行業。這些數據顯示，政府在扶植青年創業方面所扮演的角色與成果，對於國內初創企業的發展及經濟繁榮具相當程度的正向影響。

TAIDE （微調後）:

截至民國六十五年三月底為止，政府已經協助青年九百四十三人貸款創業，循環貸出新台幣一億零三百八十五萬四千元，所創事業全年生產額已達新台幣十五億餘元，創業的種類包括化工、機械製造、電機製造、運輸工具製造等二十八種行業。

可以發現未微調的 TAIDE 即使在提示詞中指定翻譯為繁體中文，原始模型仍偏好輸出摘要，這可能就是原始模型頁面上所提到的，針對自動摘要、撰寫信件、擬文章、

繁體中文翻英文、英文翻繁體中文等辦公室中常見的任務做加強 [36] 的特色，我們推測，由於原始模型輸出字數較多且對繁體中文字的支援完善，所以 TAIDE 在搭配非對稱映射機制時的效果會很好，但本身進行 zero-shot 簡繁轉換任務的效果並不及 Taiwan-LLM。此外值得注意的是，原始的 TAIDE 模型也傾向將國字的數字轉寫為阿拉伯數字。

4.5.1 消融實驗

考慮到現實層面的使用場景，接下來我們以「OpenCC 為原始翻譯」此最低成本且廣泛的情境做為基準，進一步探討各種因素對於我們實驗結果的影響。

表 4.8 是使用未微調大型語言模型做為參考的比較。No AM 代表是模型的原始 response，With AM 代表是使用了大型語言模型以及非對稱映射機制來對原始翻譯做修正進而得到的最終的翻譯結果。

表 4.8 使用與未使用非對稱映射機制的比較結果

非對稱映射機制（AM）	Chinese-Alpaca（未微調）		Taiwan-LLM（未微調）		TAIDE（未微調）	
	BLEU	GLEU	BLEU	GLEU	BLEU	GLEU
No AM	0.1439	0.2533	0.2264	0.3572	0.1370	0.2668
With AM	**0.9396**	**0.9644**	**0.9413**	**0.9654**	**0.9406**	**0.9649**

可以發現在還沒微調的情況下，無論是簡體中文或是繁體中文的大型語言模型都無法有效的完成翻譯任務，但在借助非對稱映射機制後，就可以有效的大幅提升的整體翻譯品質。這證實了第一，簡繁轉換確實具有其專業性，通用的大型語言模型無法直接泛化到此任務上。第二，此機制對於 zero-shot 的簡繁轉換而言輔助是極為巨大的，

且對無論是未微調亦或是微調後的大型語言模型而言，前二者的自然語言理解能力都較傳統語言模型好。

表 4.9 使用已微調大型語言模型做為參考的實驗結果

Model	BLEU	GLEU
Chinese-Alpaca（未微調）	0.9396	0.9644
Taiwan-LLM（未微調）	0.9413	0.9654
TAIDE（未微調）	0.9406	0.9650
Chinese-Alpaca（微調後）	**0.9590**	**0.9758**
Taiwan-LLM（微調後）	0.9455	0.9678
TAIDE（微調後）	0.9518	0.9719

表 4.9 是使用已微調大型語言模型做為參考的實驗結果。在經過微調後，以簡體中文為主要語言的 Chinese-Alpaca 效果略優於以繁體中文為主要語言的 Taiwan-LLM 以及 TAIDE，我們推測是因為這是簡轉繁的任務，因此 token 本就以簡體為主的大型語言模型自然就具有主場優勢。

值得一提的是，Taiwan-LLM 與 TAIDE 似乎有著強大的泛化效能，在三種 Baseline 的情況下都相較於 Chinese-Alpaca-2-7B 而言效果較優，且模型微調與否的翻譯效果上升幅度皆沒有後者來的大。由此可見，微調大型語言模型並搭配非對稱映射機制確實可以有效的提升本機端簡繁轉換的翻譯品質。

我們另外摘錄了幾種 Baseline 與大型語言模型的組合，Baseline 是 OpenCC，而大型語言模型則是先前實驗中效果最好的 Chinese-Alpaca 做為代表。結果如表 4.10 所示，紅色字與框框代表與正確句子不同的地方。

表 4.10 兩階段簡繁轉換框架例句

範例句的簡繁轉換和實際句子	說明
比如我说小人，不一定针对今天社会上的小人；我讲文人也不一定是讲今天我们自己。 原始翻譯：比如我說小人，不一定針對今天社會上的小人；我講文人也不一定是講今天我們自己。 模型翻譯：比如我說小人，不一定針對今天社會上的小人；我講文人也不一定是講今天我們自己。 修正後翻譯：比如我說小人，不一定針對今天社會上的小人；我講文人也不一定是講今天我們自己。 實際句子：比如我說小人，不一定針對今天社會上的小人；我講文人也不一定是講今天我們自己。	Baseline 與模型都翻譯正確的例句。
让人好奇的是，亲兄弟怎么姓氏不同？ 原始翻譯：讓人好奇的是，親兄弟怎麼姓氏不同？ 模型翻譯：讓人好奇的是，親兄弟怎么姓氏不同？ 修正後翻譯：讓人好奇的是，親兄弟怎麼姓氏不同？ 實際句子：讓人好奇的是，親兄弟怎麼姓氏不同？	Baseline 翻譯正確且模型翻譯不正確的例句。
这个成功的例子，使得许多养殖户有意效尤，相信水车普遍装设之后，虱目鱼的产量一定会更为增加。 原始翻譯：這個成功的例子，使得許多養殖戶有意效尤，相信水車	Baseline 翻譯不正確且模型翻譯正確的例句。

普遍裝設之後，蝨目魚的產量一定會更爲增加。 模型翻譯：這個成功的例子，使得許多養殖戶有意效尤，相信水車普遍裝設之後，虱目魚的產量一定會更增加。 修正後翻譯：這個成功的例子，使得許多養殖戶有意效尤，相信水車普遍裝設之後，虱目魚的產量一定會更爲增加。 實際句子：同修正後翻譯。	
在统计中显示：民国五十三、四年，妇女就业人数只有一百万人，到了民国六十至六十二年就业人数就达到一百八十三万人。在这些人中间，雇主有一千多人，自营业有二十一万九千人，协助工作者有六十一万九千人，受雇者有一百万人，其中在工厂者有八十三万人，政府机构工作者十七万人。 原始翻譯：...自營業有二十一萬九千人，協助工作者有六十一萬九千人，受僱者有一百萬人，其中在工廠者有八十三萬人，政府機構工作者十七萬人。 模型翻譯：...自營有二十萬九千人，協助工作者有六十萬九千人，受僱者有一百萬人，其中在廠子者有八十三萬人，...。 修正後翻譯：...自營業有二十一萬九千人，協助工作者有六十一萬九千人，受僱者有一百萬人，其中在工廠者有八十三萬人，...。 實際句子：同修正後翻譯。	模型翻譯正確但出現幻覺所以少或多字的例句。
近年来，中南美洲、中东、欧洲及东南亚各国的纺织工业及成衣业发展迅速，而这些地区的纽扣工业却尚未建立，对纽扣的需求极为殷切。 原始翻譯：近年來，中南美洲、中東、歐洲及東南亞各國的紡織工業及成衣業發展迅速，而這些地區的紐扣工業卻尚未建立，對紐扣的需求極爲殷切。 模型翻譯：同原始翻譯。	Baseline 翻譯與模型翻譯皆不正確的例句。

修正後翻譯：同原始翻譯。 實際句子：近年來，中南美洲、中東、歐洲及東南亞各國的紡織工業及成衣業發展迅速，而這些地區的鈕釦工業卻尚未建立，對鈕釦的需求極為殷切。	

此外，或許是因為基於《教育部重編國語辭典修訂本》的語料太過嚴謹，導致有九成以上的修正後的句子與實際用法之間都出現了超譯的現象，也就是使用了不同的字但是都可以表達出相同的意思，這些字有些是俗體字、有些是異體字有些則是早期的用法，例如「台」與「臺」、「為」與「爲」、「眾」與「衆」、「才」與「纔」等，有些甚至也沒有定論，如表 4.11 所示。

表 4.11 翻譯結果常見差異

民間常用字	翻譯結果
台	臺
為	爲
群	羣
眾	衆
啟	啓
才	纔
凶	兇
吃	喫
計畫	計劃
了解	瞭解

在我們上述的實驗中，我們並沒有對這些詞做額外的處理，期許未來能夠有更嚴謹且通用的標準看待這些早期年代與現代用法的微妙差異。

第五章 結論

本研究專注於在資源有限的環境下，改進根據語境的簡繁字對轉換的正確性。在本研究中我們首先先從教育部的《重編國語辭典修訂本》[18] 中，根據類似 Alpaca-7B 的格式進行資料清洗，並得到有 36 萬筆資料的繁體中文指令資料集 MOE-RMCD。再使用簡體中文的 Chinese-Alpaca-2-7B、以台灣文化及繁體中文見長的 Taiwan-LLM 和國科會的 TAIDE-LX-7B-Chat 這三種基於 LLaMA-2 的大型語言模型（LLM），分別進行指令遵循監督式微調（Instuction-Following Supervised Fine-Tuning），本研究發現在消費級的顯示卡上透過量化低秩適應（Quantization Low-Rank Adaptation, QLoRA）技術微調預訓練的大型語言模型，就能夠得到優於現行主流開源簡繁翻譯軟體的結果。

但是本研究也觀察到直接由大型語言模型所輸出的結果會出現輸入與輸出句子不一致，也就是「幻覺」（Hallucination）的問題。因此進一步的提出了非對稱映射機制（Asymmetric Mapping），這是一種巧妙利用簡體與繁體之間所存在的非對稱關係的規則性映射，它會將原始翻譯內的非對稱簡繁字強制轉換為大型語言模型所轉換的非對稱簡繁字，這樣既可以使用到大型語言模型對於上下文的理解，又可以確保最終結果不會有少字或是多字的情況發生。

從實驗中我們可以得知，第一，簡繁轉換確實具有其專業性，通用的大型語言模型無法直接泛化到此任務上。第二，非對稱映射機制對於 zero-shot 的簡繁轉換而言輔助是極為巨大的，能夠快速提升其效果，第三，無論是未微調亦或是微調後的大型語言模型，前二者的自然語言上下文理解能力都較傳統語言模型好。

關於未來可能的工作包含像是加入對俗體字、異體字或是早期用法的共通特性；對指令做提示工程（prompt engineering）；將此非對稱映射機制應用於不同參數規模的大型語言模型上，例如 13B 或甚至 70B 的大型語言模型；改變修正幻覺的方法，例如使用檢索強化生成（RAG）；以及期望能夠更妥善的利用此繁體中文指令資料集 MOE-RMCD 等等。

參考文獻

[1] 中研院資訊所 （2023），"中研院資訊所對 CKIP-Llama-2-7b 之回應"，https://www.sinica.edu.tw/News_Content/70/1850,（accessed on December 7, 2023）.

[2] OpenAI （2022）, “Introducing ChatGPT”, https://openai.com/blog/chatgpt, （accessed on December 29, 2023）.

[3] Agarwal, A., Wong-Fannjiang, C., Sussillo, D., Lee, K., & Firat, O. (2018). Hallucinations in Neural Machine Translation., ICLR

[4] 中華人民共和國文字改革委員會（1964），《簡化字總表》。

[5] 馬英九 （2004），"《也是「正名運動」—為「繁體字」正名為「正體字」請命》"，https://archive.ph/20120714132234/http://theme.taipei.gov.tw/cgi-bin/SM_theme?page=424aab17,（accessed on December 18, 2023）.

[6] 中華民國經濟部（2007），《商品檢驗法施行細則》第九條與第二十一條。

[7] 中華民國大陸委員會（2013），《大陸地區物品勞務服務在臺灣地區從事廣告活動管理辦法》第四條。

[8] 中華民國衛生福利部（2014），《藥物製造業者檢查辦法》第十六條。

[9] 中華民國教育部（1982），《常用國字標準字體表》。

[10] 中華民國教育部（2011），《標準字與簡化字對照手冊》。

[11] Li, M. H., Wu, S. H., Zeng, Y. C., Yang, P. C., & Ku, T. (2010). Chinese characters conversion system based on lookup table and language model. Computational Linguistics and Chinese Language Processing, 15(1), 19-36.

[12] Chen, Y., Shi, X., & Zhou, C. (2011, November). A simplified-traditional chinese character conversion model based on log-linear models. In 2011 International Conference on Asian Language Processing (pp. 3-6). IEEE.

[13] Shi, X., Chen, Y., Huang, X. （2013）. Key problems in conversion from simplified to traditional Chinese characters. Paper presented at the Proceedings of the XIV Machine Translation Summit, Nice, Italy.

[14] Hao, T., & Zhu, C. (2011, June). Simplified-traditional Chinese character conversion based on multi-data resources: Towards a fused conversion algorithm. In The 2nd International Conference on Next Generation Information Technology (pp. 50-56). IEEE.

[15] Xu, J., Ma, X., Tsai, C. T., & Hovy, E. (2017, November). STCP: Simplified-Traditional Chinese Conversion and Proofreading. In Proceedings of the IJCNLP 2017, System Demonstrations (pp. 61-64).

[16] Hao, T., & Zhu, C. （2013）. Toward a professional platform for chinese character conversion. ACM Transactions on Asian Language Information Processing, 12（1）, Article 1.

[17] Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). OPTQ: Accurate quantization for generative pre-trained transformers. In The Eleventh International Conference on Learning Representations.

[18] 中華民國教育部（1994），《重編國語辭典修訂本》。

[19] Halpern, J., & Kerman, J. (1999). The pitfalls and complexities of Chinese to Chinese conversion. In MTSummit (pp. 458-466).

[20] Heafield, K. (2011). KenLM: Faster and smaller language model queries. In Proceedings of the sixth workshop on statistical machine translation (pp. 187-197).

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[22] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. （2023）. Alpaca: A Strong, Replicable Instruction-Following Model. Stanford Center for Research on Foundation Models. https://crfm.stanford.edu/2023/03/13/alpaca.html.

[23] Unicode Consortium. （2021）. CJK Unified Ideographs, from https://www.unicode.org/charts/PDF/U4E00.pdf（accessed on January 15, 2024）

[24] 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. arXiv preprint arXiv:2106.09685.

[25] Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A., Lester, B., Du, N., Dai, A., & Le, Q.（2021）. Finetuned Language Models Are Zero-Shot Learners. ArXiv, abs/2109.01652.

[26] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[27] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[28] Lin, Y. T., & Chen, Y. N. (2023). Taiwan llm: Bridging the linguistic divide with a culturally aligned language model., CoRR, abs/2311.17487

[29] Cui, Y., Yang, Z., & Yao, X. (2023). Efficient and effective text encoding for chinese llama and alpaca. arXiv preprint arXiv:2304.08177.

[30] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems, 36.

[31] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[32] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[33] Open Chinese Convert 開放中文轉換，https://github.com/BYVoid/OpenCC （accessed on April 4, 2024）.

[34] Google 翻譯，https://translate.google.com.tw/ （accessed on April 4, 2024）.

[35] TAIDE，https://taide.tw/index （accessed on May 24, 2024）.

[36] TAIDE-LX-7B 與 TAIDE-LX-7B-Chat 模型，https://huggingface.co/taide/TAIDE-LX-7B-Chat （accessed on May 24, 2024）.