Rewriting Chinese Educational Materials to Change Readability Levels with Large Language Models: Strategies and Challenges (利用大型語言模型改寫中文教育文本可讀性：策略與挑戰)

Hojin Koh[1], Wan-Jun Gu[1], Hou-Chiang Tseng[1], Kuan-Yu Chen[1], Yao-Ting Sung[2]
[1] National Taiwan University of Technology
[2] National Taiwan Normal University
hojinkoh@citrinefox.com wjgu@nlp.csie.ntust.edu.tw
tsenghc@mail.ntust.edu.tw kychen@mail.ntust.edu.tw sungtc@ntnu.edu.tw

摘要

歷來教材改寫在教育領域是重要卻十分耗時費力的工作，本研究探討利用開源大型語言模型改寫中文教育文本以調整文本可讀性。使用的方法為無需進行模型微調的零樣本(Zero-shot) 提示方法。大型語言模型在英文文本改寫任務中表現優異，但對中文文本效果有限。為改善中文文本改寫效果，本研究提出並評估了三種策略：跨語言改寫、具體年級目標提示和迭代改寫。這些策略在一定程度上提高了中文教育材料的改寫效果，但仍存在挑戰。本文探討了大型語言模型在教育材料改寫中的優勢與局限性，並討論了潛在的教育應用前景。

Abstract

Rewriting educational material has long been an important but labor-intensive task. This paper explores the potential of using open-source large language models (LLMs) to rewrite Chinese educational materials for different grade-level readability. The main methodology is zero-shot prompting approaches without any fine-tuning. While LLMs demonstrated excellent performance in rewriting English materials, their effectiveness for Chinese materials was limited. To address this, we proposed and evaluated three strategies: cross-lingual rewriting, specific grade-level targeting, and iterative rewriting. Our findings suggest that these strategies improved the rewriting of Chinese educational materials, though challenges remain. We provide an in-depth analysis of the advantages and limitations of LLMs in educational material rewriting and discuss potential educational applications.

關鍵字：中文文本可讀性、英文文本可讀性、大型語言模型、改寫、零樣本學習

Keywords: Chinese Readability, English Readability, Large Language Models, Rewriting, Zero-shot Learning

1 緒論

可讀性 (Readability) 長久以來一直是教育研究中適性閱讀一個重要的基石。可讀性是指閱讀材料能夠被讀者所理解的程度 (Dale and Chall, 1949; Klare, 2000; Mc Laughlin, 1969; De Clercq and Hoste, 2016)。當學習者的程度和學習材料的可讀性相符合時，會產生較好的理解及記憶效果 (Klare, 2000)，若是太難或是太簡單則都會干擾學習 (Cambria and Guthrie, 2010)。由於文件的可讀性在知識傳遞扮演極為重要的角色，因此研究人員持續發展各種能夠自動且精準地估測文本可讀性的模型。這些模型從早期使用計算語言學的特徵來研發出各式可讀性公式及語言特徵 (Chall and Dale, 1995; Feng et al., 2010; Hong et al., 2016)，到近期採用表示學習法 (Representation Learning) 改善分類效果 (Tseng et al., 2016) 或是採用深度類神經網路 (Deep Neural Networks) 進行分類等方法 (Tseng et al., 2017)，都致力於提高可讀性評估的準確性。

然而，僅僅能夠評估文本的可讀性並不足以解決教育實踐中的問題。一部份的挑戰在於如何有效地改寫改寫文本的難度以應用於課堂之上，使其適應不同年級學生的閱讀能力。這種改寫不僅要調整文本的難度，還需要保持原始內容的核心資訊和教育價值。

成功的文本改寫能帶來諸多好處：首先，它能夠確保學生接觸到難度適中的學習材料，從而提高學習效率和動機；其次，它能

夠幫助教師更靈活地使用各種教學資源，適應不同學生的需求，因為就算是同年級的學生，每一個人的閱讀能力也不盡相同。然而，有效改寫教材是十分耗時費力的工作，不僅需要熟悉不同年級適合的閱讀難度，也要確保改寫後的教材仍然能有效傳達知識。

近年來，隨著自然語言處理技術的快速發展，大型語言模型(Large Language Models)在多個領域展現出強大的潛力。這些模型能夠生成非常自然的人類語言，並有限度地理解人類提供的指示，為教育材料的自動改寫開闢了新的可能性。特別是開源的大型語言模型，為研究人員提供了靈活且強大的工具，可以用於探索如何自動調整教育材料的難度，以適應不同年級學生的需求。此外，大型語言模型的多語言能力 (Armengol-Estapé, 2021; Lai et al., 2023) 為跨語言的教育材料改寫提供了可能性，對英文教學上的幫助是有潛力的。

本研究旨在探討如何有效地運用大型語言模型，通過各種提示方法，來改寫不同學科的教育材料，以適應較低年級或較高年級的適讀程度。如此不僅能夠提供一種自動化調整教育材料難度的新方法，還能深入分析大型語言模型在這一應用場景中的優勢與局限性，為未來的教育科技研究提供有價值的見解。

2 文獻回顧

可讀性的應用範疇不僅限於教育材料，還廣泛涉及法律文件和醫療資訊等領域，並聚焦於如何評估和改善各類文本的可讀性 (Collins-Thompson, 2014)。例如，有學者探討了法律文件的可讀性問題，試圖透過可讀性分析來提高法律文本的易讀性 (Curtotti et al., 2015)。同樣地，在醫療領域，研究人員致力於簡化醫療資訊，以確保患者能夠準確理解重要的健康資訊 (Zheng et al., 2022)。

隨著自然語言處理技術的進步，特別是生成式語言模型，如 GPT 系列模型 (Radford et al., 2018; 2019) 的出現，研究人員開始嘗試利用這些技術來改寫複雜文本。早期的一些研究主要集中在如摘要及資訊檢索系統中查詢語句的改寫 (Zhang et al., 2019a; Chen et al., 2020) 等應用。這些研究雖然主要目的並非改變文本的難度，但從廣義上來說，這些工作實際上也在一定程度上改變了文本的複雜度和可理解性，因此可以被視為是一種廣義的可讀性調整。這些研究為後續的教育材料改寫奠定了重要基礎。

近年來，隨著大型語言模型的革命性發展，有研究者開始直接將大型語言模型應用於教育材料的可讀性調整。在英語學習材料方面，Huang et al. (2024) 亦嘗試使用大型語言模型將文本改寫至特定的難度等級 (Lexile Level)。這些研究不僅探討了大型語言模型在教育材料改寫中的潛力，還為如何評估和控制改寫後文本的可讀性提供了新的思路。

3 研究方法

3.1 資料集

本研究使用了兩種語言的教育材料作為實驗資料集：中文和英文。

中文資料集選自 98 年度臺灣翰林、康軒、南一三大出版社所出版的一年級到十二年級審定版教科書，涵蓋國語科、社會科、自然科及體育和健康教育等領域。這些教科書均經由專家根據課程綱要編製而成，確保內容對該年級學生的適切性。

英文資料集則涵蓋了台灣一至十二年級的英語教科書文本及作為十三年級文本的高三升大學英文科考題，包含翰林、康軒、南一、何嘉仁等出版社。其中除了十三年級的考題外，一至十二年級的文本皆為主課文。

中文文本		英文文本	
年級	文本數	年級	文本數
1 年級	149	1 年級	56
2 年級	192	2 年級	56
3 年級	334	3 年級	45
4 年級	356	4 年級	44
5 年級	370	5 年級	44
6 年級	363	6 年級	42
7 年級	676	7 年級	102
8 年級	707	8 年級	108
9 年級	595	9 年級	84
10 年級	831	10 年級	196
11 年級	866	11 年級	193
12 年級	789	12 年級	177
		12 以上	290

表 1. 中英文實驗文本在各年級的數量分佈

特徵	維度
句數、字/詞數、獨特字/詞數、獨特字/詞比例	7
每句平均字/詞數、每句最大字/詞數	4
字平均/最大筆劃數、10 劃以下/10-20 劃/20 劃以上字比例	5
平均/最大詞長、一字詞/二字詞/三字詞/四字以上詞比例	6

表 2. 中文可讀性模型之語言特徵

中文文本資料集共計 6,228 篇文本，英文文本資料集共計 1,437 篇文本，兩種語言的文本在各年級的數量分佈如表 1 所示。

3.2 文本可讀性模型

為了評估改寫後文本的難度，本研究針對中文文本和英文文本各採用一個預先訓練好的可讀性模型。此模型能夠準確地將文本分類為一年級到十二年級之間的適讀程度。

對於中文文本，本研究的可讀性模型採用了多維度的特徵表示方法，結合了詞向量技術和傳統語言學特徵，並使用支援向量機 (Support Vector Machine, SVM) 進行分類 (Vapnik and Chervonenkis, 1974)。具體來說，對每一則文本，本研究使用了每一個單詞 250 維的 GloVe 詞向量 (Global Vectors for Word Representation) (Pennington et al., 2014) 之平均向量和 22 維的額外語言學特徵 (Liu et al., 2015)，共 272 維特徵，語言特徵詳細組成如表 2 所示。GloVe 向量使用的訓練資料為中文維基百科，能夠有效捕捉中文文本中的語義資訊。

可讀性模型的分類效果在本研究中參照 Tseng et al. (2017) 等過去研究以鄰近正確率 (Adjacency Accuracy) 來評估。中文文本若將資料隨機切成五份進行 5 折交叉驗證 (5-fold Cross-validation)，難度分類鄰近正確率為 85.44%；若盡可能將不同出版社放在不同分割中，在等分的前提下進行對抗式分割 (Adversarial Split: Søgaard et al., 2021)，則 5 折交叉驗證的鄰近正確率為 80.03%。

對於英文文本，本研究使用支援向量機搭配 300 維的 GloVe 詞向量之平均向量進行分類。GloVe 向量取自史丹佛大學預訓練的 GloVe 向量，使用 Common Crawl 資料訓練，840B tokens 及 2.2M 詞彙量的版本。英文可讀性模型 5 折交叉驗證的分類鄰近正確率為 83.62%。

這兩個可讀性模型為評估大型語言模型改寫文本的效果提供了重要的基準和評估工具。

3.3 開源大型語言模型

由於本研究所使用的資料有版權上的考量，不適合使用如 GPT-4 等須透過網路傳輸資料的商業大型語言模型。本研究採用 Meta 公司於 2024 年 4 月發布可離線運作的 Llama3-Instruct 模型，包含 8b 和 70b 兩種參數量的變體（後續陳述實驗結果時，此二模型將簡稱為 8b 和 70b）。Llama3 是 Meta 基於 Llama 系列的架構 (Touvron et al., 2023) 推出的新一代開源大型語言模型。

雖然此模型相較於其他模型擴增了詞彙表，使其在多國語言上表現較好[1]，但其訓練資料仍以英文為大宗，應在英文的文本改寫任務能夠表現較好 (Armengol-Estapé, 2021)；此模型上下文長度則為 8192 個 token，可適用於許多不同長度的文本改寫任務。本研究中對大型語言模型所使用的提示語詳見文末的圖 9 和圖 10。

在接下來的實驗中，本研究將詳細探討 Llama3-Instruct-8b 和 Llama3-Instruct-70b 在文本改寫任務中的表現，及其在不同語言間的適應能力。

3.4 評估準則

為了評估大型語言模型在教材文本改寫任務中的表現，本研究採用以下評估指標：

1. 方向性準確度 (Directional Accuracy)：使用前述之可讀性模型衡量模型是否能夠按照指示將文本改寫為更容易閱讀或更難閱讀的版本。
2. 難度變化 (Readability Level Change, RLC)：使用前述之可讀性模型測量改

1 https://ai.meta.com/blog/meta-llama-3/

寫前後文本難度的平均變化程度，了解模型調整文本難度的幅度。

3. 語義相似度 (Semantic Similarity)：使用 BERTScore (Zhang et al., 2019b) 計算原始文本和改寫後文本的語義相似度，確保改寫過程中不至於和原文的核心含義差異過大。

3.5 初步實驗觀察

為了探討大型語言模型在教育本文改寫任務中的表現，本研究進行了一系列初步的實驗。這些實驗涵蓋了兩種不同參數規模的模型（前述之 8b 和 70b），兩種改寫方向（改簡單和改難），以及兩種語言的資料集（英文和中文）。實驗結果呈現出顯著的語言差異，為後續研究提供了方向的指引。

英文文本的改寫任務初步方向性準確度實驗結果如圖 1 所示，兩顆模型都展現出優異的表現。無論是提高還是降低文本的難度，模型都能夠有效地調整文本難度。尤其是 70b 模型表現更是十分亮眼，僅經過一次的改寫便可使文本平均變化最多 2.30 個難度等級；以 BERTScore 評估改寫前後相似性，其 F1 Score 最多可高達 0.922，也就是改寫後的文本在很大程度上保留了原始文本的語義內容，說明模型能夠在調整難度的同時，維持文本的核心資訊。

然而，當面對中文文本時，兩種模型的表現都有明顯的下降。雖然以 BERTScore 評估相似性顯示語義上仍得到相對完整的保留；但大部分文本在改寫後仍然被判定為原有的難度，如圖 2 所呈現的方向性準確度結果所示，只有約 20% 的文檔呈現與提示語指示相符的難度變化，平均最多只能使文本平均變化 0.77 個難度等級。此初步實驗與後續實驗的難度變化和語義相似度的資訊可參考表 3。

這些初步的實驗觀察突顯了大型語言模型在處理較不擅長語言時的表現差異，本研究接著將在實驗設計中嘗試設計弭平此差異的策略，以改善應用在中文文本上的效果。

3.6 實驗設計

從初步實驗觀察可以發現大型語言模型在英文文本改寫任務中表現優異，但在中文文本上的效果有限。為了改善模型在中文文本改寫方面的表現，本研究設計了以下三種改寫策略：

1. 跨語言改寫策略：考量大型語言模型在英文文本改寫任務中的優異表現，以及這類模型在機器翻譯方面的強大能力 (Zhang et al., 2023)，本研究提出基於翻譯的跨語言改寫策略：首先將原始中文文本使用大型語言模型翻譯成英文，接著使用同樣的大型語言模型對英文文本進行可讀性改寫，最後將改寫後的英文文本翻譯回中文。這樣子的策略期望能夠充分利用模型在英文文本處理方面的優勢，同時保持中文文本的語義完整性。
2. 具體年級目標提示策略： 在初步實驗中，提示語僅要求模型將文本改寫得「更容易」或「更難」理解。為了提供更明確的指引，本研究提出在提示語中加入具體的年級目標，例如「請改寫成符合小學三年級學生的閱讀能力」或「請改寫成符合高中三年級學生的閱讀能力」。這種方法的出發點在於具體的年級目標可能會激發模型對不同教育階段語言特徵更具體的理解 (Zamfirescu-Pereira et al., 2023)，從而產生更有效的改寫效果。
3. 迭代改寫策略： 儘管初步實驗中中文文本的整體改寫效果不佳，但部分文本仍然有一定程度的難度變化。基於這一觀察，本研究提出一種迭代改寫策略：對每個文檔進行最多 5 次的連續改寫，每次都使用前一次改寫的結果作為輸入。這種方法的假設是，通過累積多次小幅度的改變，最終較有可能可以達到預期的難度調整效果 (Madaam et al., 2024)。

4 實驗結果與分析

4.1 跨語言改寫策略的中文文本改寫效果

為了改善大型語言模型在中文文本改寫任務中的表現，本研究首先嘗試了將原始中文文本翻譯成英文，接著以英文進行可讀性改寫，最後再將改寫後的英文文本翻譯回中文的跨語言改寫策略。由於模型在英文文本

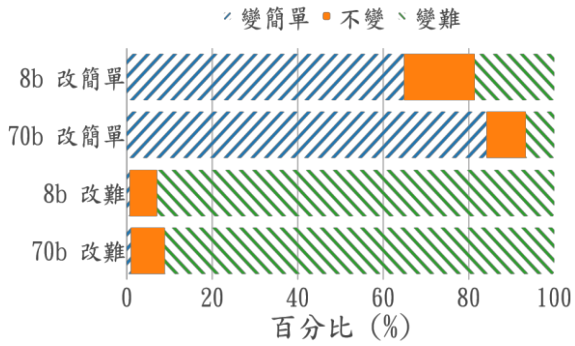

圖 1. 英文文本改寫效果

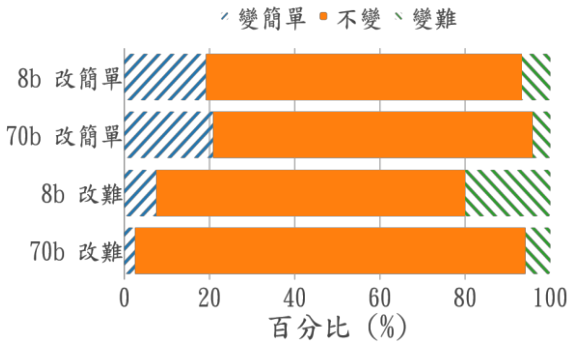

圖 2. 中文文本初步改寫效果

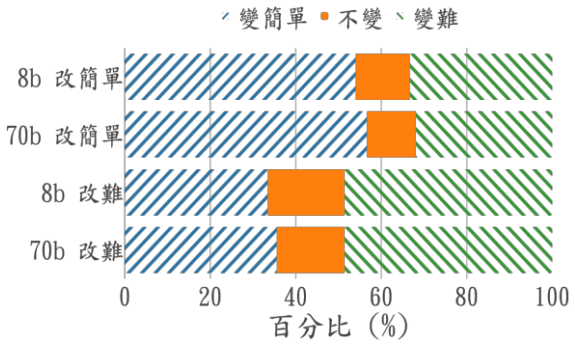

圖 3. 跨語言改寫策略的中文文本改寫效果

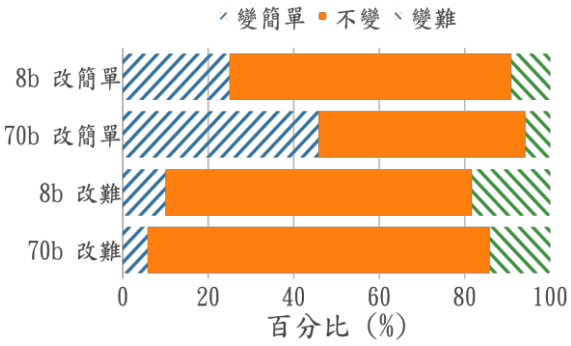

圖 4. 具體年級目標提示策略的中文文本改寫效果

改寫方面表現優異，這種方法應會產生較理想的改寫效果。

然而，實驗結果顯示，跨語言改寫策略雖然有一定的效果——純粹就方向性準確度而言，確實有較好的表現——但結果並不如預期穩定。本組實驗的方向性準確度結果如圖 3 所示，無論是使用 70b 還是 8b 模型、改寫為更簡單還是更難的版本，文本難度變化的方向都呈現出相當大的不穩定性：許多文本被改寫成了與預期相反的方向。

這種不穩定性造成了不理想的平均難度變化，四組實驗中最好的一組（70b 改簡單）文本平均僅變化 0.80 個難度等級，與中文文本的比較基準相比差距不大，並仍遠低於英文文本改寫的效果。此外，改寫後文本與原文的語義相似度也有明顯下降，改寫前後 BERTScore 的 F1 Score 介於 0.611（8b 改難）和 0.698（70b 改簡單）之間。

因此，雖然跨語言改寫策略在某些情況下可以提高方向性準確度，但多次翻譯過程會使原始的語義資訊逐漸流失，且中文與英文間可能缺乏一致的難度理解，而使得難度在翻譯的過程中產生變化，造成改寫方向不穩定。

儘管如此，跨語言改寫策略的部分成功仍然提供了有價值的見解。它證明了利用模型在某種語言上的優勢來改善其在另一種語言上的表現是可行但不穩定的，在未來的後續研究中也許可以嘗試針對此不穩定性進行改善。

4.2 具體年級目標提示策略的中文文本改寫效果

在探索了跨語言策略後，本研究轉向了另一種中文文本改寫的改進方法：在提示中加入具體的年級目標。這種策略的目的是為模型提供更明確的指導，希望能激發模型對不同教育階段語言難度深層的理解。

這種策略在使用較大規模的 70b 模型時取得了一定程度的改善，本組實驗的方向性準確度結果如圖 4 所示。相較於中文文本的比較基準，具體年級目標提示策略明顯提高了方向性準確度，亦沒有出現跨語言改寫策略中觀察到的不穩定現象。

在難度變化方面，此組實驗中也觀察到了較大的改善。使用具體年級目標提示策略後，平均難度變化最高增加到了 1.10 個難度等級，這表明模型能夠更有效地調整文本的難度。同時，改寫前後 BERTScore 的 F1 Score

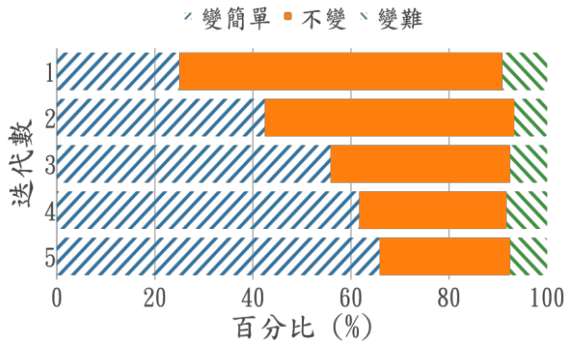

圖 5. 迭代改寫策略：8b 改簡單

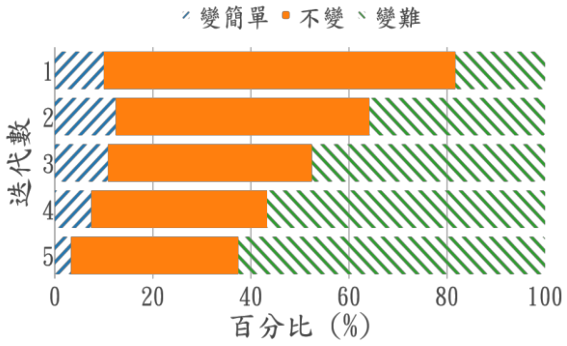

圖 6. 迭代改寫策略：8b 改難

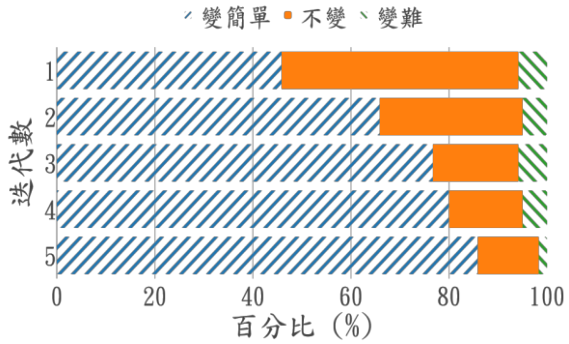

圖 7. 迭代改寫策略：70b 改簡單

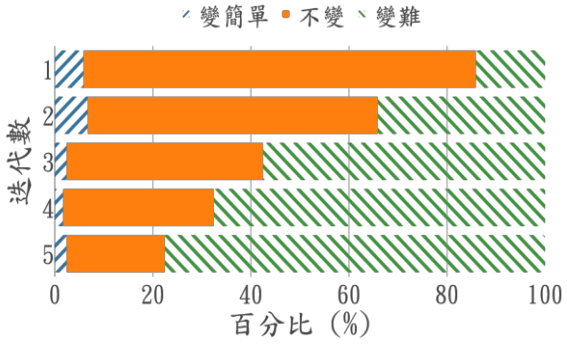

圖 8. 迭代改寫策略：70b 改難

與初步實驗的結果相近，顯示語義相似度與比較基準相比大致保持不變。

然而，值得注意的是，儘管有了這些改進，中文文本改寫的表現仍然未能達到英文文本改寫的水準。這可能反映了模型在處理中文時的固有局限性，或者是中文文本在難度調整上的特殊挑戰；另外使用較小規模的 8b 模型時，具體年級目標提示策略並未帶來明顯的改善，也就是這種較複雜的提示方式，較小規模的大型語言模型可能無法有效處理。這也暗示對於資源受限無法使用較大的大型語言模型的使用情境，可能需要探索其他更有效的策略。

4.3 迭代改寫策略的中文文本改寫效果

在確認了具體年級目標提示策略的有效性後，本研究建基於該策略之上，進一步探討迭代改寫策略對中文文本改寫的影響。這種策略的核心思想是通過多次連續改寫，每次都以前一次改寫的結果作為當次改寫的原始文本，藉此累積小幅度的變化，最終加大能夠進行可讀性改寫的輻度。

本實驗對前項策略的四種實驗組合分別進行了最多 5 次的迭代改寫。本組實驗的方向性準確度結果如圖 5 到圖 8 所示，其中僅迭代一次的改寫實質上就是前一節「具體年級目標提示策略」的結果。隨著迭代次數的增加，所有實驗組合都呈現出明顯的改善趨勢。這表示即使原始文本已經是同一個模型使用同樣提示語改寫後的結果，模型仍然能夠持續維持文本難度調整的方向。特別是在使用 70b 模型時，這種效果更為顯著。

其次，在難度變化方面，迭代改寫策略的效果尤為突出：在多次迭代後，70b 改簡單和 70b 改難這兩組實驗的平均難度變化甚至超過了英文文本改寫的效果。也就是說，即使模型本身較擅長處理英文文本，透過適當的策略與投注較多的運算資源，可以在中文文本改寫任務中達到與英文相當甚至更好的效果。

然而，本研究中也觀察到，隨著迭代次數的增加，改寫後文本與原文的語義相似度略有下降：如 70b 改難第 1 次到第 5 次迭代 BERTScore 的 F1 Score 從 0.856 降到 0.771，顯示每一次改寫都會損失原始文本中一小部份的資訊。儘管如此，考慮到方向性準確度和難度變化的顯著提升，此程度的語義相似度下降是一個可以接受的權衡。

迭代改寫策略為改善中文文本的可讀性調整提供了一種有效的方法，尤其是在使用較大規模模型（如 70b）時。這種策略不僅能

策略	平均難度等級變化 (年級)				平均語義相似度 (F1 Score)			
	改簡單		改難		改簡單		改難	
	8b	70b	8b	70b	8b	70b	8b	70b
初步實驗 (英文)	-1.18	-1.21	+2.30	+1.99	0.916	0.922	0.915	0.922
初步實驗 (中文)	-0.82	-0.77	+1.01	+0.80	0.858	0.826	0.866	0.870
中文跨語言策略	-0.78	-0.80	+0.09	+0.61	0.627	0.698	0.611	0.665
中文具體年級策略	-0.85	-1.10	+0.99	+1.06	0.855	0.819	0.847	0.856
中文迭代策略 (迭代 5)	-1.55	-1.96	+2.48	+2.50	0.677	0.730	0.702	0.771

表 3. 各改寫策略的平均平均難度變化與平均語義相似度

夠顯著提高改寫的準確性和效果，還為克服大型語言模型在中文處理上的固有挑戰提供了一種可行的解決方案。

4.4 綜合分析與觀察

在三種策略中，跨語言改寫策略雖然在某些情況下能提高方向性準確度，但其不穩定性和語義保留問題限制了其實際應用。相較之下，具體年級目標提示策略和迭代改寫策略都展現出了良好的效果。特別是迭代改寫策略，不僅顯著提升了方向性準確度和難度變化的幅度，在某些情況下甚至超越了英文文本改寫的效果。這一結果令人鼓舞，表明通過適當的策略設計，可以一定程度地克服現有以英文為主的大型語言模型在處理中文時的固有局限性。

在不同的策略實驗中，結果一致顯示相較於改寫為更難的版本，將中文文本改寫為較簡單的版本效果較好；也就是在需要簡化複雜教材以適應低年級學生需求的情況下，可以有較好的表現。

當使用的策略有較多步驟時，隨著步驟量的增加，雖然文本難度的變化較顯著，但改寫後文本與原文的語義相似度會降低。在實際應用中，需要在改寫效果和保留原意之間找到適當的平衡點。

本研究亦發現模型規模對改寫效果有顯著影響。在大多數情況下，70b 模型的表現優於 8b 模型。這突顯了模型參數量在處理複雜語言任務中的重要性。而對於資源受限無法使用較大模型的情況，通過迭代改寫策略，即使使用較小規模的模型，也能在多次迭代後達到一定的成效。

5 潛在限制與模型相關衍生問題

儘管本研究在利用大型語言模型改寫中文文本方面取得了一定的成果，但仍存在一些潛在的限制和相關的衍生問題值得關注。首要的挑戰之一是硬體需求。即使使用較小規模的模型，運行大型語言模型仍然需要相當高的硬體配置，特別是需要性能良好並擁有足夠記憶體的顯示卡。這可能會限制本研究方法在缺乏必要的硬體設備時的應用。

其次，本研究採用零樣本(Zero-shot) 提示方法來控制模型輸出，但這種方法對精確調整模型的輸出較為困難。不同模型對相同的提示有不同的反應，這種不確定性可能會影響改寫結果的穩定性和可靠性，且較難確定應用在其他大型語言模型上時的效果。

此外，在實驗過程中，我們發現基於對話的大型語言模型容易在輸出中添加額外的註釋（甚至時常是以英文加注的註釋），即使在提示語中被要求不要這樣做。為了獲得純粹的改寫內容，本研究需要額外的後處理步驟作為實作細節來過濾這些非內容片段，增加了整個改寫過程的複雜性。

最後，大型語言模型也可能存在潛在的偏見，特別是在處理不同文化背景的教育內容時。確保改寫後的文本在文化上的適當性和公平性是一個需要持續關注的問題。同時，雖然本研究主要關注可讀性的調整，但在實際教育場景中，確保改寫後內容的教育準確性有時更重要，因此本研究的方法較適合在有專家介入的情況下作為一個有力的輔助工具。

6 結論與未來展望

本研究探討了利用大型語言模型改寫中文文本以適應不同年級閱讀能力的可能性。透過系列實驗和策略探索，本研究發現大型語言模型在英文文本改寫任務中表現優異，但在中文文本上的效果相對有限，突顯了跨語言應用的挑戰性。

為改善中文文本改寫的效果，本研究提出並評估了三種策略：跨語言改寫、具體年級目標提示和迭代改寫。結果顯示，後兩種策略特別是迭代改寫策略能顯著提升改寫的準確性和效果，甚至在某些情況下超越英文文本改寫。這表明通過適當的策略設計，透過大型語言模型與可讀性模型之間的相互合作，可以克服一部份現有大型語言模型在處理中文時的局限性。此外，本研究也發現模型規模和改寫目標（簡化或增難）對改寫效果有顯著影響。

儘管取得了一定進展，但現有方法仍面臨一些局限性，如硬體需求、模型輸出控制難度、後處理需求等。未來研究可以嘗試針對中文文本進行專門的模型微調，以進一步提升性能並使改寫的輸出較為穩定。同時，將教育領域的專業評估方法更緊密地整合到改寫過程中，也是一個值得探索的方向，以確保改寫後的文本不僅可讀性適當，亦能維持原有教育上的價值。

總結而論，本研究為利用大型語言模型調整各類中文文本的可讀性開闢了新的可能性。隨著技術進步和更多研究的投入，相信這種方法將為客製化教材和自適應學習材料的發展帶來重要貢獻，最終造福更廣泛的學習者群體。

Acknowledgement

This work was supported by the National Science and Technology Council of Taiwan under Grants NSTC 113-2410-H-011 -001, NSTC 112-

系統提示語	改簡單提示語	改難提示語
你是一位人工智慧教材編輯。你用台灣用語的繁體中文依照使用者的需求將指定的教材改寫成較難或較易懂的寫法。避免使用其他語言。請避免使用英文。請不要在回應中包含表情符號。你的回應裡只可包含改寫後的文章，不可加上任何前言、補述及說明文字。若改寫後的文本有長篇的英文，請將其翻譯成台灣用語的繁體中文。	以上是一篇待改寫的原始教材。在這項任務中，我們需要將此文本改寫成適合較低年級的學生閱讀的教材，同時保持原始的含義和資訊。請協助將此教材文本改寫得較容易閱讀、較適合低年級學生的閱讀理解能力、運用較少抽象概念、並使用較簡單的比喻和詞彙，且避免修辭和複雜句構的使用。	以上是一篇待改寫的原始教材。在這項任務中，我們需要將此文本改寫成適合較高年級的學生閱讀的教材，同時保持原始的含義和資訊。請協助將此教材文本改寫得較難閱讀、較適合高年級學生的閱讀理解能力、運用較多涉及抽象概念的描述、並使用較困難的比喻、詞彙、修辭和句構。

圖 9. 中文文本改寫所使用的提示語

系統提示語	改簡單提示語	改難提示語
You are an editor of AI instructional materials. Using English, rewrite the specified instructional materials according to the user's request into a simpler or easier-to-understand format. Avoid using other languages. Do not include emojis in your response. Your response should only contain the rewritten text without any preambles, additional explanations, or annotations.	The above is an original instructional material that needs to be rewritten. In this task, we need to rewrite the text to make it suitable for lower grade students to read, while preserving the original meaning and information. Please help to revise this instructional material to make it easier to read, suitable for lower grade students' reading comprehension ability, using fewer abstract concepts, and employing simpler analogies, words, and idioms, and avoiding complex sentence structures.	The above is an original instructional material that needs to be rewritten. In this task, we need to rewrite the text to make it suitable for higher grade students to read, while preserving the original meaning and information. Please help to revise this instructional material to make it more challenging to read, suitable for higher grade students' reading comprehension ability, using more abstract concepts and descriptions, and employing more difficult analogies, words, idioms, and sentence structures.

圖 10. 英文文本改寫所使用的提示語

2628-E-011-008-MY3 and NSTC 113-2640-B-002-005. This project was financially supported by the "Empower Vocational Education Research Center" of the National Taiwan University of Science and Technology (NTUST) from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. We thank the National Center for High-performance Computing of the National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

References

Armengol-Estapé, Jordi, Bonet, Ona de Gibert, and Melero, Maite, 2021. On the multilingual capabilities of very large-scale English language models. *arXiv preprint* arXiv:2108.13349.

Cambria, Jenna and Guthrie, John T., 2010. Motivating and engaging students in reading. *New England Reading Association Journal*, *46*(1), pp.16-29.

Chall, Jeanne S. and Dale, Edgar, 1995. *Readability revisited: The new Dale-Chall readability formula.* Cambridge, Mass: Brookline Books.

Chen, Zheng, Fan, Xing, and Ling, Yuan, 2020. Pre-training for query rewriting in a spoken language understanding system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pages. 7969-7973.

Collins-Thompson, Kevyn, 2014. Computational assessment of text readability: A survey of current and future research. ITL-International Journal of Applied Linguistics, 165(2), pages 97-135.

Curtotti, Michael, McCreath, Eric, Bruce, Tom, Frug, Sara, Weibel, Wayne, and Ceynowa, Nicolas, 2015, June. Machine learning for readability of legislative sentences. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law,* pages 53-62.

Dale, Edgar and Chall, Jeanne S., 1949. The concept of readability. *Elementary English*, *26*(1), pages 19-26.

De Clercq, Orphée and Hoste, Véronique, 2016. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, *42*(3), pages 457-490.

Feng, Lijun, Jansche, Martin, Huenerfauth, Matt, and Elhadad, Noémie, 2010. A comparison of features for automatic readability assessment. *Coling 2010: Posters* pages 276-284.

Hong, Jia-Fei, Sung, Yao-Ting, Tseng, Ho-Chiang, Chang, Kuo-En, and Chen, Ju-Ling, 2016. A multilevel analysis of the linguistic features affecting Chinese text readability. *Taiwan Journal of Chinese as a Second Language*, (13), pages 95-126.

Huang, Chieh-Yang, Wei, Jing, and Huang, Ting-Hao'Kenneth, 2024. Generating Educational Materials with Different Levels of Readability using LLMs. *In2Writing 2024,* arXiv:2406.12787.

Klare, George R., 2000. The measurement of readability: useful information for communicators. *ACM Journal of Computer Documentation (JCD)*, *24*(3), pages 107-121.

Lai, Viet Dac, Ngo, Nghia Trung, Veyseh, Amir Pouran Ben, Man, Hieu, Dernoncourt, Franck, Bui, Trung, and Nguyen, Thien Huu, 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint* arXiv:2304.05613.

Liu, Yi-Nian, Chen, Kuan-Yu, Tseng, Ho-Chiang and Chen, Berlin, 2015. A study of readability prediction on elementary and secondary Chinese textbooks and excellent extracurricular reading materials. In *Proceedings of the 27th Conference on Computational Linguistics and Speech Processing (ROCLING 2015)* pages 71-86.

Madaan, Aman, Tandon, Niket, Gupta, Prakhar, Hallinan, Skyler, Gao, Luyu, Wiegreffe, Sarah, Alon, Uri et al., 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, *36*.

Mc Laughlin, G. Harry., 1969. SMOG grading-a new readability formula. *Journal of reading*, *12*(8), pages 639-646.

Radford, Alec, Narasimhan, Karthik, Salimans, Tim, and Sutskever, Ilya, 2018. Improving language understanding by generative pre-training.

Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, and Sutskever, Ilya, I., 2019. Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), page 9.

Pennington, Jeffrey, Socher, Richard, and D. Manning, Christopher, 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),* pages 1532-1543.

Søgaard, Anders, Ebert, Sebastian, Bastings, Jasmijn, and Filippova, Katja, 2021. We Need To Talk About Random Splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823-1832.

Touvron, Hugo, Lavril, Thibaut, Izacard, Gautier, Martinet, Xavier, Lachaux ,Marie-Anne, Lacroix, Timothée, Rozière, Baptiste, et al., 2023. Llama:

Open and efficient foundation language models. *arXiv preprint* arXiv:2302.13971.

Tseng, Hou-Chiang, Sung, Yao-Ting, Chen, Berlin, and Lee, Wei-En, 2016. Classification of text readability based on representation learning techniques. In *Proceedings of the 26th Annual Meeting of the Society for Text & Discourse*, pages 1-6.

Tseng, Hou-Chiang, Chen, Berlin, and Sung, Yao-Ting, 2017. Exploring the use of neural network based features for text readability classification. *International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP)*, *22*, pages 31-46.

Vapnik, Vladimir Naumovich and Chervonenkis, Alexey Yakovlevich, 1974. *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya* (Theory of pattern recognition. Statistical problems of learning). Moscow, Russia: Nauka.

Zamfirescu-Pereira, J. D., Wong, Richmond, Hartmann, Bjoern, and Yang, Qian., 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems,* pages 1-21.

Zhang, Haoyu, Cai, Jingjing, Xu, Jianjun, and Wang, Ji, 2019. Pretraining-Based Natural Language Generation for Text Summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* pages 789-797.

Zhang, Biao, Haddow, Barry, and Birch, Alexandra, 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning,* pages 41092-41110.

Zhang, Tianyi, Kishore, Varsha, Wu, Felix, Weinberger, Kilian Q., and Artzi, Yoav, 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zheng, Yifan, Tang, Yan, Tseng, Hou-Chiang, Chang, Tao-Hsing, Li, Lanping, Chen, Pan, Tang, Yubo, Lin, Xiao-bin, Chen, Xiao, and Tang, Ke-Jing, 2022. Evaluation of quality and readability of over-the-counter medication package inserts. *Research in Social and Administrative Pharmacy*, *18*(9), pages 3560-3567.