Proceedings of the 36th Conference on Computational Linguistics and Speech Processing (ROCLING 2024)

Organized by
Institute of Linguists, Academia Sinica
Research Center for Information Technology Innovation, Academia Sinica
Institute of Information Science, Academia Sinica
The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

November 4th–5th
Academia Sinica, Taipei City, Taiwan

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

Published by The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

Shu-Chuan Tseng, Yu Tsao, Hen-Hsen Huang, Yao-Chung Fan, and Chia-Hui Chang (eds.)

Proceedings of the 36th Conference on Computational Linguistics and Speech Processing (ROCLING 2024)

2024-11-04–2024-11-05
ACLCLP
2024-11

ISBN 978-986-95769-7-0

Preface by the General Co-Chairs of ROCLING 2024

It is our great pleasure to welcome you to the 36th Conference on Computational Linguistics and Speech Processing (ROCLING 2024), the annual conference of the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). For more than three decades, ROCLING has been a vibrant gathering place for researchers and practitioners in natural language processing and speech processing, advancing both fundamental theories and real-world applications.

This year, ROCLING returns to Academia Sinica in Taipei, Taiwan, from November 4 to 5, 2024. Academia Sinica has had the honor of hosting this conference on several occasions in the past—in 1989, 1993, 1997, 2004, and 2017—and we are delighted to welcome you back once again. It is truly special for us to continue this tradition of bringing together our community in this historic venue.

The program for ROCLING 2024 is rich and diverse. We are especially honored to feature keynote speeches by two internationally renowned scholars: Prof. Eduard Hovy (Carnegie Mellon University) and Prof. Heng Ji (University of Illinois Urbana-Champaign). Their contributions to computational linguistics and artificial intelligence have inspired researchers worldwide, and we are privileged to have them share their insights with us this year. In addition, the conference will showcase high-quality oral papers rigorously selected through peer review, as well as presentations of the TAIDE project. Together, these elements highlight both the depth of foundational research and the breadth of practical innovation in our field.

In recent years, the field has been transformed by the rise of large language models (LLMs). Their ability to engage in fluent dialogue, provide responsive interactions, and demonstrate comprehension has captured the imagination of both the research community and the public at large. Against this dynamic backdrop, ROCLING 2024 offers a timely opportunity for experts, scholars, and practitioners to explore the evolving landscape of computational linguistics and speech technologies, and to exchange ideas that will shape the next chapter of our field.

We would like to express our heartfelt gratitude to Program Chair Yao-Chung Fan (National Chung Hsing University) and Keynote Chair Chia-Hui Chang (National Central University) for their invaluable leadership and dedication. We are also deeply grateful to the program committee, reviewers, and the entire organizing team for their tireless efforts, as well as to all the authors and participants whose contributions make this conference possible.

We extend our sincere appreciation to our sponsors, whose generous support has been instrumental in making ROCLING 2024 a reality: Tagtoo Co., Ltd., Institute of Information Science (Academia Sinica), Research Center for Information Technology Innovation (Academia Sinica), Industrial Technology Research Institute, Delta Electronics, eLAND Information Co., Ltd., Chunghwa Telecom, Cyberon Corporation, and the TAIDE Project.

Most importantly, we thank you, the members of our community, for bringing your passion, curiosity, and energy to ROCLING 2024.

General Co-Chairs of ROCLING 2024
Shu-Chuan Tseng, Academia Sinica, Taipei, Taiwan
Yu Tsao, Academia Sinica, Taipei, Taiwan
Hen-Hsen Huang, Academia Sinica, Taipei, Taiwan

Organizing Committee

General Co-chairs
Shu-Chuan Tseng, Academia Sinica, Taipei, Taiwan
Yu Tsao, Academia Sinica, Taipei, Taiwan
Hen-Hsen Huang, Academia Sinica, Taipei, Taiwan

Program Committee Chair
Yao-Chung Fan, National Chung Hsing University, Taichung, Taiwan

Keynote Chair
Chia-Hui Chang, National Central University, Taoyuan, Taiwan

Program Committee
Kai-Wei Chang
Yung-Chun Chang
Chien-Hung Chen
Chung-Chi Chen
Kuan-Yu Chen
Po-Chun Chen
Yi-Pei Chen
Chen-Yu Chiang
Min-Yuh Day
Yao-Chung Fan
Hen-Hsen Huang
Lun-Wei Ku
Lung-Hao Lee
Yu-Ting Lee
Wei-Yun Ma
Ming-Hsiang Su
Ming-Feng Tsai
Richard Tzong-Han Tsai
Yu Tsao
Hou-Chiang Tseng
Hsin-Min Wang
Jenq-Haur Wang
Sheng-Lun Wei
Jheng-Long Wu
An-Zi Yen

Contents

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

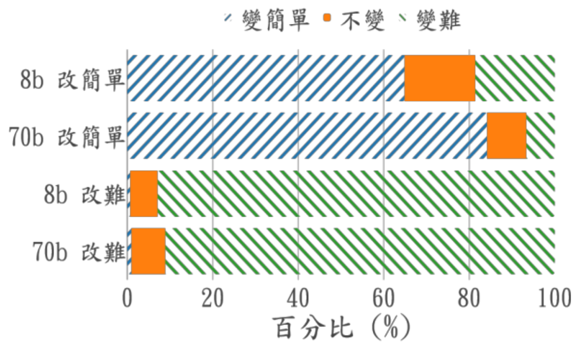


圖 1. 英文文本改寫效果

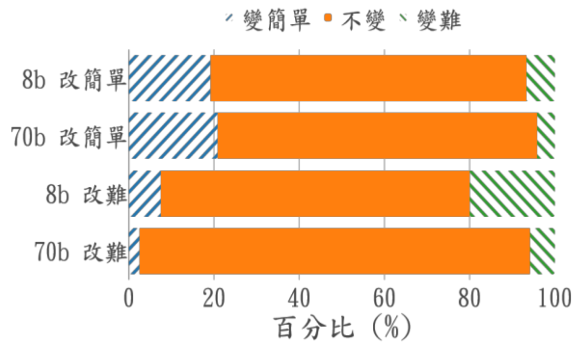


圖 2. 中文文本初步改寫效果

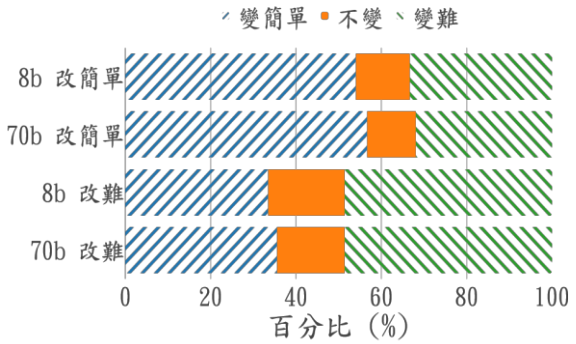


圖 3. 跨語言改寫策略的中文文本改寫效果

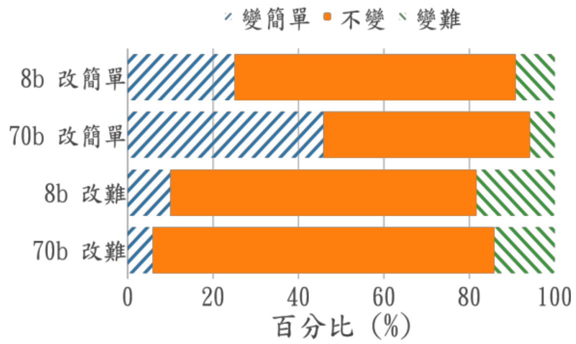


圖 4. 具體年級目標提示策略的中文文本改寫效果

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

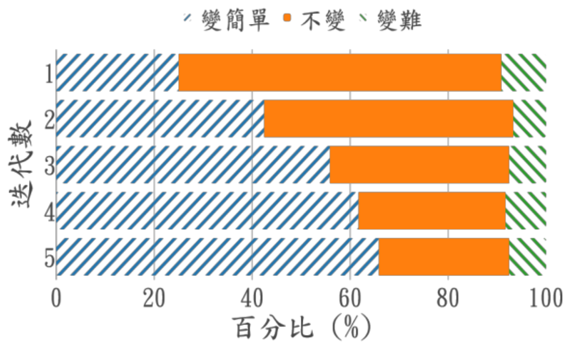


圖 5. 迭代改寫策略：8b 改簡單

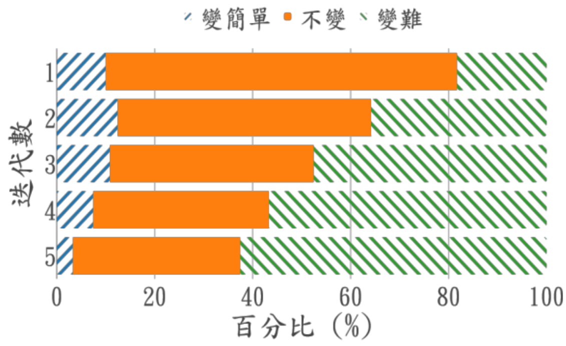


圖 6. 迭代改寫策略：8b 改難

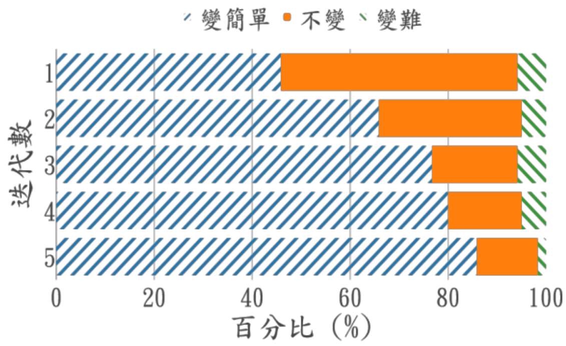


圖 7. 迭代改寫策略：70b 改簡單

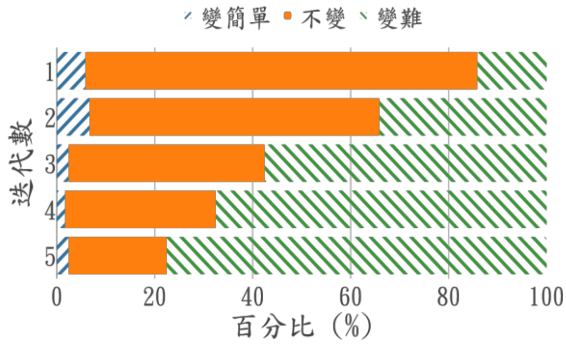


圖 8. 迭代改寫策略：70b 改難

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幅度。

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.

Quantity Affects Quality: Instruction Fine-Tuning on LLM's Multiple-choice Question Abilities

Hsuan-Lei Shao
Graduate Institute of Health and Biotechnology Law
Taipei Medical University
hlshao@tmu.edu.tw

Wei-Hsin Wang
Sieh-Chuen Huang
College of Law
National Taiwan University
wangweihsin07@gmail.com
schhuang@ntu.edu.tw

Abstract

This paper discovered the potential of instruction fine-tuning to significantly performance of large language models (LLMs) on legal multiple-choice questions (MCQs) abilities. By manipulating the volume of training data, we aim to demonstrate a strong correlation between the quantity of data used in fine-tuning can lift LLM's quality, paving the way for LLMs in specific task (ex: legal knowledge). We compared Breeze-7B (based on Mistral-7B) and its fine-tuned version. Adding more MCQs data can enhance their abilities, there are two models: the first is adding 5,000 new samples(bz5k), and the second is 70,000(bz70k). We compare these with the general baseline model, GPT-3.5, GPT-4o, and one traditional Mandarin LLM(TAME). Then, the MCQs dataset of the MMLU, TMMLU, and the 2023 Taiwanese Bar Examination be evaluated. We find that fine-tuning LLMs might degrade its original capabilities little. However, surpassing a specific data volume can markedly enhances the model's effectiveness. This balance ensures that while the LLM's proficiency in specialized legal domains is enhanced. Practically speaking, we developed a legal MCQ-specific LLM that demonstrated the benefits of model customization. For specialized applications, smaller-scale, personalized LLMs can be developed with reduced training costs, making advanced legal tools more accessible and adaptable to specific knowledge areas or unique legal frameworks. This approach also addresses concerns about digital sovereignty by aligning the model's functionalities with jurisdiction-specific legal regulations.

1 Introduction

1.1 The Evolution of Legal Informatics: Large Language Models in Legal Contexts

Legal informatics, an interdisciplinary field that merges legal studies with information technology, has evolved significantly since its inception. It began primarily with the automation of legal documentation and progressed to more complex applications, including data management and electronic access to statutes and case law. This development was spurred by the legal profession's need to manage increasing volumes of information and the desire for more efficient legal processes[1, 2]

In the early days, legal informatics focused on creating databases for case law and legislation, facilitating quicker and more reliable access to legal resources. As technology advanced, the field expanded to include tools for legal analysis, document automation, and even predictive technologies that could forecast litigation outcomes[3, 4]. Furthermore, visualization techniques have become an important methodological step in translating legal texts into formal languages, bridging the gap between human understanding and machine processing [5].

Recent advancements in artificial intelligence have further propelled legal informatics towards innovation. The collaboration between legal, computational, and data science communities aims to build innovative legal models to improve the existing legal system[6]. Comprehensive overviews of legal informatics, such as the work by Katz and Dolin [7], provide valuable insights into real-world applications like document review and online dispute resolution.

1.2 Challenges in Implementing LLMs Across Diverse Legal Systems

The integration of Large Language Models (LLMs) in legal practices has shown promising results in areas such as document drafting and legal research. However, applying general LLMs faces significant challenges in specialized fields like legal regulation, where every country may have distinct laws and regulations. This specificity requires LLMs to understand and adapt to diverse legal frameworks, a task that general models are currently ill-equipped to handle[8, 9].

The risk of losing legal diversity is significant, particularly for smaller countries or unique cultural contexts. These regions often have legal nuances that are not well represented in the vast data pools used to train standard LLMs. This phenomenon, known as "sovereignty AI," highlights the need for models that respect and incorporate different jurisdictions' legal sovereignty and specificities [10, 11]. To address this, there is a growing push for developing customized LLMs that are trained on localized data sets, ensuring that the legal advice and documentation generated are relevant and compliant with local laws[12, 13].

We utilized a “Mandarin version” derivative of the Mistral-7B model, named Breeze-7B, which was further enhanced through prompt finetuning. This process involved the integration of an additional set of MCQs and answers aimed at improving the model's capabilities in legal contexts. Then, we developed two variations of this model: one finetuned with 5,000 new samples (bz5k) and another with 70,000 new questions (bz70k). Our findings reveal a nuanced interplay specializing legal knowledge in the performance of LLMs. Specifically, the models Breeze-7B, bz5k and bz70k showed that finetuning with insufficient data volumes can indeed degrade the model's original capabilities, negatively impacting the architecture designed for knowledge tasks. Conversely, when the data volume surpassed a certain threshold (as with bz70k), the model's effectiveness significantly improved in legal knowledge but can influent other tack performance.

2 Literature Review

2.1 Overview of LLM Capabilities in Legal Domains

As we mentioned, LLMs have increasingly become integral to various applications within legal domains, demonstrating capabilities that span from basic legal information retrieval to complex reasoning and document generation. Studies have shown that LLMs, like the GPT series and its successors, can interpret, generate, and summarize legal texts with a high degree of accuracy. These models have been employed for contract analysis, litigation prediction, and even in assisting with legal education by generating hypothetical legal scenarios for study. This section reviews the extent of LLM integration in legal practices and evaluates their effectiveness in handling diverse legal tasks [8, 15, 19].

2.2 Current Methodologies in Instruction Fine-Tuning

Instruction fine-tuning is a recent development aimed at refining the training process of LLMs to better follow user instructions. Unlike traditional model training, instruction fine-tuning focuses on aligning the model's outputs with specific user expectations and requirements. In the legal field, this is particularly advantageous for ensuring that models adhere to legal reasoning patterns and comply with jurisdiction-specific regulations. This segment will cover the latest methodologies in instruction fine-tuning, including the application of specialized datasets (like legal judgments or statutory provisions) that train models to recognize and replicate the nuanced decision-making processes typical in legal analyses[23,24].

3 Research Design

3.1 Multiple Choice Questions in Legal Evaluation

We use a basic method in legal LLMs–multiple choice questions (MCQs). The MSQ plays a critical role in legal education and professional assessments. Moreover, the structured nature of MCQs makes them particularly suitable for automation using AI technologies like LLMs. By incorporating LLMs in creating and grading MCQs, educational institutions can enhance the objectivity and efficiency of assessments. LLMs can also be used to generate diverse question sets that cover a wide array of topics, providing a robust tool for comprehensive legal training[14, 15].

However, the effectiveness of LLMs in this area depends heavily on their training and the quality of data used. It is essential that the data reflects the specific legal principles and practices relevant to the jurisdiction where the education or assessment is taking place. This ensures that the questions are accurate and contextually appropriate, fostering a more effective and meaningful learning environment[16, 17].

3.2 Model Instructure Finetuning: Breeze-7B

In this study, we based the capabilities of the Breeze-7B-base model[27], which is built upon the foundations of the Mistral-7B architecture, by incorporating an extensive set of MCQs into its training regimen.

The original Breeze model, without any specific finetuning towards these datasets, serves as a control to understand the baseline capabilities of the LLM. The bz5k model, finetuned with 5,000 samples, represents a modest increase in dataset-specific training. The bz70k model, representing a substantial fine-tuning effort with 70,000 samples, aims to tailor the model towards the dataset characteristics significantly.

Our approach to instruction fine-tuning involves directly using the multiple-choice options as inputs. The output consists of the correct option (A, B, C, or D) along with the content of the option, which provides additional information. This method ensures that the model selects the correct answer and understands the context and details associated with each option, enhancing its ability to handle similar questions effectively. For example:

```
{
    "input": "Question: Which of the following is NOT considered in assessing capacity for liability? (A) The ability to recognize that an action is illegal (B) The ability to control one's actions (C) The mental state at the time of the action (D) The ability to choose between legal and illegal actions",
    "output": "(B) The ability to control one's actions "
},
{
    "input": "Question: After A grants B a permit for hillside development and B transfers the land to C, does the original permit still apply to C? (A) Yes (B) No (C) Depends on the situation (D) None of the above",
    "output": "(A) Yes "
},
```

Figure 1: Instructure Finetuning Datasets Structure

3.3 Evaluation Design

These models were benchmarked against a general baseline model, GPT-3.5, the more advanced GPT-4, and a traditional Mandarin Large Language Model (TAME)[28]. The datasets employed for evaluation included the Multimodal Legal Understanding (MMLU)[29], the Taiwanese Multimodal Legal Understanding (TMMLU)[30], and the 2023 Taiwanese Bar Examination questions[18].

The evaluation of these models was conducted using two distinct methodologies to assess their performance in legal multiple-choice question scenarios:

- Probability Selection Method for MMLU Dataset: This method involves the extraction of probabilities corresponding to the multiple-choice options 'A', 'B', 'C', and 'D' using a specific function designed to interact with the LLM's output layers. The option with the highest probability is selected as the model's response. This approach is feasible with available LLM configurations where computational costs are within manageable limits[20] . In practical, we used the code provide by the MMLU dataset directly.

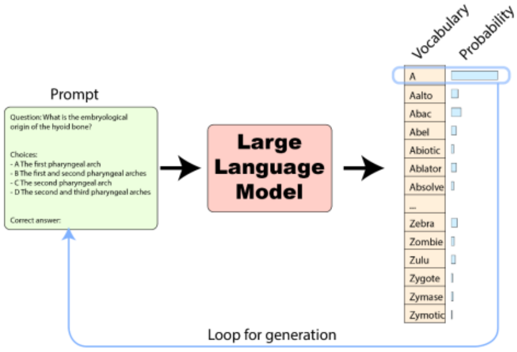


Figure 2: Probability Selection Method[25]

- Prompt-Based Zero-Shot Evaluation: The second evaluation method utilizes a zero-shot approach through direct prompting, which can be applied not only to our custom models but also in conjunction with external services such as the OpenAI API. This enables us to include and assess the performance of other models like GPT-3.5 and GPT-4 in a straightforward and practical manner, leveraging their built-in capabilities without additional fine-tuning[21].

Our approach to evaluating prompts involves a straightforward method where we directly input the question, for example,*"Question: After A grants B a permit for hillside development and B transfers the land to C, does the original permit still apply to C? (A) Yes (B) No (C) Depends on the situation (D) None of the above"*. We then expect it to respond with "(A) Yes". Therefore, we extract the first option that appears in the output (A, B, C, or D). If none of these options (A, B, C, or D) appear in the output, we default to "C" for consistency across all models.

Given the availability of correct answers, we are able to calculate the performance of all models on MCQs. In this study, we employ "accuracy" as the criterion to assess the efficacy of each model. This metric allows us to quantitatively evaluate how well the models are performing in selecting the correct responses from the provided options[22].

Table 1: Comparing Different Finetuning Quantity Effects

Dataset\Model	Breeze	bz5k	bz70k
TMMLU(Law)	0.407	0.401	**0.486**
TMMLU(Enginnering)	**0.498**	0.493	0.458
MMLU	0.560	**0.562**	0.515
TBE	0.486	0.457	**0.514**

note:TBE = “the 2023 Taiwanese Bar Examination”

4 Research Result and Discussion I: Finetuning Quantity Effect

4.1 Probability Selection Evaluation

The table "Comparing Different Finetuning Quantity Effects" showcases the impact of varying quantities of data used in finetuning on the performance of the Breeze model across different datasets. These datasets encompass the Taiwanese Multi-Modal Legal Understanding (TMMLU) in Law and Engineering domains, the broader Multi-Modal Legal Understanding (MMLU), and the 2023 Taiwanese Bar Examination (TBE).

1. Dataset-Specific Performance: TMMLU (Law) and TMMLU (Engineering): For the Law subset of TMMLU, increasing the finetuning quantity results in improved performance, as evident from the bz70k model’s score of 0.486 compared to the bz5k’s 0.401 and the baseline’s 0.407. This suggests that a larger dataset helps the model better understand and adapt to legal nuances.

   Conversely, in the Engineering subset, the performance decreases as the quantity of finetuning increases (0.458 in bz70k down from 0.498 in the baseline). This could indicate overfitting or perhaps the introduction of noise or less relevant information through the additional data.

2. Different Language Performance: MMLU

   Here, we see a slight improvement in bz5k over the baseline (0.562 vs. 0.560), but a reduction with bz70k (0.515). This pattern suggests that while some targeted finetuning can be beneficial, excessive finetuning may lead to diminishing returns or negative transfer, where too much specificity detracts from the model’s general applicability. This demonstrates that if a LLM performs better in one language, it often performs worse in another. This may be related to the parameters of individual tokens, where finetuning can detrimentally affect the original linguistic structure of the LLM.

3. The Newest Local Knowledge: TBE

   Performance on the Taiwanese Bar Examination dataset improves significantly with the highest data volume (bz70k), moving from 0.486 to 0.514. This improvement indicates that comprehensive legal training data can enhance model performance on specialized legal tasks such as bar exams, which likely benefit from a deeper understanding of localized legal principles and practices.

4.2 Discussion

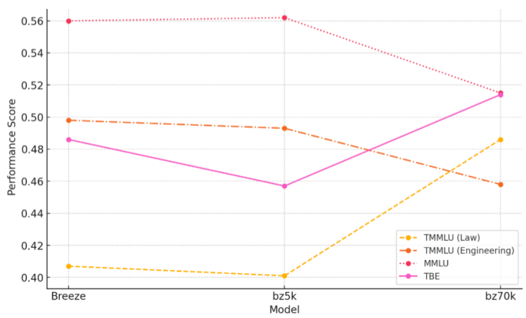


Figure 3: Model Quantity with Each Evaluation Dataset

The graph represents the performance comparison of three models (Breeze, bz5k, bz70k) across four different datasets (TMMLU-Law, TMMLU-Engineering, MMLU, TBE), with each model serving as a point on the x-axis and performance scores on the y-axis. Different line styles distinguish each dataset.

1. TMMLU (Law) (dotted line): Shows a trend of improvement as the finetuning data volume increases, peaking with the bz70k model.

2. TMMLU (Engineering) (dash-dot line): This line trends downward, indicating a decrease in performance with more extensive finetuning, potentially due to overfitting or less relevant finetuning data for engineering-specific content.

3. MMLU (dashed line): Performance slightly increases with moderate finetuning (bz5k) but decreases with extensive finetuning (bz70k),

suggesting that a balance needs to be found to avoid diminishing returns.

4. TBE (solid line): Shows a recovery in performance with the most extensive finetuning (bz70k), indicating that larger, more focused datasets may be beneficial for specialized legal examinations like the bar exam.

This graph visually illustrates how varying the amount of finetuning data impacts model performance across different domains. It highlights the need for careful consideration of how much and what type of data to use for finetuning to optimize performance without compromising the model's generalization capabilities. This insight is crucial for applying LLMs in specialized fields where accuracy and specificity are paramount.

5 Research Result and Discussion II: Prompt Evaluation

5.1 Prompt Evaluation

In the second phase, we input the MCQs string by the API directly (refer to 3.3. Evaluation Design), which allows us to pull other outside LLMs to compare.

Table 2: Comparing Different Finetuning Quantity Effects by Prompting Input

Dataset\Model	Breeze	bz5k	bz70k	TAME	GPT-3.5	GPT-4o
TMMLU(administrative_law)	0.250	0.380	**0.580**	0.480	0.336	0.650
MMLU	0.320	0.540	0.590	0.470	**0.660**	0.860
TBE	0.106	0.423	**0.640**	0.390	0.423	0.680

The table provides comparative performance data for different language models on MCQs across three datasets: TMMLU (administrative law, subcategory of the Law category), MMLU, and TBE. It shows how each model fares in accurately responding to prompts within these specific domains.

First of all, we wish to skip the discussion on GPT-4o because its performance is too strong, making it only possible for us to attempt to approach its performance; moreover, because its size is much larger, it is not comparable to our approximately 7B parameter model. We can see the trends on the Breeze-series and other LLMs:

1. Impact of Finetuning Method: Because our instruct fine-tuning itself has enough MCQs diversity, the bz70k model can achieve high performance when we ask it directly. It happens on TAME's performance, which is on other fields of TMMLU performances are better than bz70k. Only because it was not "familiar" with the instructions.

2. Quantity Affects Quality: This table clearly illustrates that finetuning with a larger volume of data specifically tailored to the task at hand can significantly enhance a model's performance. The bz70k's success across datasets indicates that the additional specific training it received is highly effective than Breeze-base and bz5k, even the GPT-3.5.

3. General vs. Specialized Models: The comparison between bz70k variants and GPT-4.0 highlights an essential aspect of language model application: general models can perform well across broad tasks, but under domain-specific fine-tuning processing, smaller LLM (7B) can reach the larger performance.

5.2 Discussion

The graph illustrates the performance comparison across different language models on three datasets: TMMLU (administrative law), MMLU, and TBE (Taiwanese Bar Examination). Each model is represented on the X-axis, and the performance score, likely accuracy or a similar metric, is represented on the Y-axis. Different line styles and colors distinguish the performance of each dataset.

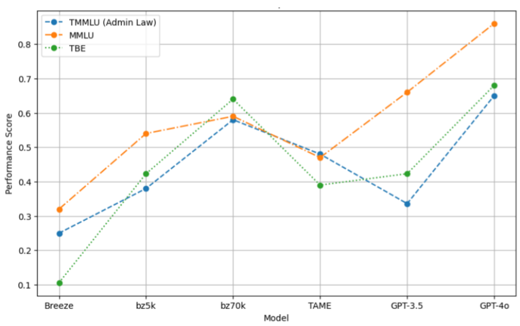


Figure 4: More Model Performance with Each Evaluation Dataset

Trend Analysis of the graph:

1. Incremental Improvements: The graph illustrates a clear trend of incremental performance improvements as we move from the baseline Breeze model to the bz5k and then to the bz70k. This trend is evident across all datasets but varies in magnitude.

2. TMMLU (Admin Law): For the TMMLU (Admin Law) dataset, the performance improvement from Breeze (0.25) to bz5k (0.38) and then to bz70k (0.58) is quite pronounced. This significant uptick suggests that the additional training samples used in finetuning the bz70k model are highly effective at enhancing the model's capabilities in handling complex administrative law scenarios.

3. MMLU: The trend in the MMLU dataset follows a similar pattern. Starting from a performance score of 0.32 with Breeze, there is a noticeable increase to 0.54 with bz5k, and further improvement to 0.59 with bz70k. This consistent increase across finetuning stages underscores the effectiveness of using larger, more targeted training sets for enhancing model performance in general legal contexts.

4. TBE: In the TBE dataset, the performance jumps considerably from Breeze (0.106) to bz5k (0.423), and sees a significant peak at bz70k (0.64). This demonstrates that extensive finetuning with a large volume of specialized data is particularly beneficial for models that navigate the complexities of bar examination questions, which likely involve nuanced legal reasoning and specific legal knowledge.

6 Conclusion

6.1 Finetuned Small LLMs Can Battle

The fine-tuned trend observed from Breeze through bz5k to bz70k highlights the significant role that the volume and specificity of finetuning data play in enhancing model performance across diverse legal datasets. We particularly emphasize the representativeness and practicality of the Taiwanese Bar Examination (TBE) dataset as a significant representative of local knowledge, underscoring its value for testing the efficacy of AI models in handling specific legal contexts relevant to Taiwan. This aligns with the broader need in AI development for datasets deeply embedded in particular legal and cultural environments, thus serving as practical tools for evaluating how well AI technologies can adapt to localized conditions.

The success of bz70k in accurately handling the TBE dataset indicates that with sufficient targeted training, LLMs can achieve high levels of proficiency in legal reasoning and analysis. This is promising for deploying AI in legal practices, where accuracy, understanding of local laws, and practical applicability are paramount. This progression supports the effectiveness of incremental finetuning strategies and emphasizes the necessity of aligning model training with the specific demands of the tasks and datasets to optimize performance in specialized applications like law.

In other words, the central assertion of the text is that the quantity of data, particularly when it is of high quality, plays a crucial role in enhancing the performance of AI models. This principle is reflected in the paper's title, "Quantity Affects Quality," which posits that substantial inputs in terms of data can translate into significant improvements in the output capabilities of a model, even if the initial performance of the model is not particularly strong. While the quantity of the finetune data is highlighted as a key factor, the quality of this data is equally important. High-quality data for finetuning ensures that the model learns relevant and accurate information, which is crucial for effectively applying the model in real-world scenarios.

Our improvement plan has achieved the original research targets:

1. Localized and Personalized LLMs: Digital Sovereignty and Localization: The project successfully integrates localized legal knowledge into LLMs through finetuning processes. This approach aligns with the growing demand for digital sovereignty, where regions or organizations dictate the informational and operational contours of the technology they deploy. Personalizing LLMs to reflect local legal standards and knowledge bases enhances their practicality and relevance, ensuring that the generated content and advice are legally sound and contextually appropriate.

2. Data Volume and Model Performance: The research confirms that significant enhancements in model performance can be achieved by increasing the volume of training data used during finetuning. This finding is crucial for smaller models, which might not start out with the computational or architectural advantages of larger models like GPT-3.5 or GPT-4.0. By effectively using larger datasets for finetuning, these smaller models can bridge the performance gap, challenging the notion that bigger is always better. The acknowledgment that future advancements in base models could lever-

age similar finetuning strategies underscores the iterative nature of AI development. This forward-looking perspective encourages continuous adaptation and enhancement as newer and more robust technologies emerge.

3. Practical Application and Instructional Method: Alignment with Use Scenarios: By employing direct input of questions as instructions, the project aligns model usage with real-world application scenarios, particularly in legal practices. This method improves the practical usability of the LLM, as it mimics the actual inquiries and tasks that users would perform, thus providing more accurate and contextually relevant responses.

6.2 Research Limitations and Future Prospects:

1. Technological and Resource Limitations: The limitation in experimenting with larger models due to equipment constraints is a common challenge in computational research. Larger models typically require substantial computational resources, which may not be readily available to all research institutions. Seeking partnerships and support can facilitate access to more advanced computational resources, enabling the exploration of larger, more capable models that may offer enhanced performance and new capabilities. The pursuit of collaborative efforts with institutions that have the necessary infrastructure can accelerate research and development efforts, pooling resources for mutual benefit.

2. Expanding the Scope of Legal Informatics: While MCQs are a common format for testing and training AI systems due to their structured nature, legal reasoning represents a more complex challenge that involves understanding nuances and making judgments similar to those a human lawyer would make. Expanding LLM research to include these aspects can significantly impact the legal profession by providing tools that can assist with more sophisticated tasks. Legal Reasoning and AI: Future research could focus on enhancing the capabilities of LLMs to handle complex legal reasoning and argumentation, potentially revolutionizing how legal analysis and advice are delivered.

3. Improving Evaluation Methods: Current evaluation methods might not adequately capture the universality and reproducibility necessary for legal applications. Legal AI systems must produce consistent and reliable results under various conditions to be truly effective and trustworthy. Therefore, developing more robust evaluation frameworks that can more accurately assess the effectiveness of AI in legal contexts is essential. This might involve creating standardized datasets, developing new metrics that better reflect the quality of legal reasoning, or adopting more rigorous cross-validation techniques to ensure the AI's decisions are sound and defensible.

The concept that "Quantity Affects Quality" underlines the transformative potential of data volume and quality in AI development. It suggests that strategic finetuning with carefully selected data can significantly uplift even underperforming models, broadening the scope for AI enhancements and applications across various industries. This principle not only guides practical AI development strategies but also sets a foundation for future research into effective and efficient AI training methodologies.

References

[1] Jenkins J. What can information technology do for law. Harvard J Law Technol. 2008;21:589.

[2] Waltl B, Zec M, Matthes F. A data science environment for legal texts. In: Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL). 2015. p. 193-194.

[3] Conrad JG, Al-Kofahi K, Zhao Y, Karypis G. Effective document clustering for large heterogeneous law firm collections. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2005. p. 177-187.

[4] Pacheco HR, Pavez MM. Contemporary challenges in legal informatics: Workshop INJU. In: 2016 11th Iberian Conference on Information Systems and Technologies (CISTI). 2016. p. 1-2.

[5] Lachmayer F, Cyras V. Visualization of legal informatics. J Vis Law. 2021;3:3-10.

[6] Sharma S, Gamoura S, Prasad DM, Aneja A. Emerging legal informatics towards legal innovation: Current status and future challenges and opportunities. Legal Inf Manage. 2021;21:218-235.

[7] Katz D, Dolin R. Legal informatics. Cambridge: Cambridge University Press; 2021.

[8] Šavelka J, Ashley KD. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. Front Artif Intell. 2023;6:1279794. doi:10.3389/frai.2023.1279794.

[9] Sun Z. A short survey of viewing large language models in legal aspect. ArXiv. 2023;abs/2303.09136. doi:10.48550/arXiv.2303.09136.

[10] Shaghaghian S, Feng L, Jafarpour B, Pogrebnyakov N. Customizing contextualized language models for legal document reviews. In: 2020 IEEE International Conference on Big Data (Big Data). 2020. p. 2139-2148. doi:10.1109/BigData50022.2020.9378201.

[11] Zhang D, Petrova A, Trautmann D, Schilder F. Unleashing the power of large language models for legal applications. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 2023. doi:10.1145/3583780.3615993.

[12] Trozze A, Davies TP, Kleinberg B. Large language models in cryptocurrency securities cases: Can ChatGPT replace lawyers? ArXiv. 2023;abs/2308.06032. doi:10.48550/arXiv.2308.06032.

[13] Elwany E, Moore DA, Oberoi G. BERT goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding. ArXiv. 2019;abs/1911.00473.

[14] Shui R, Cao Y, Wang X, Chua T. A comprehensive evaluation of large language models on legal judgment prediction. ArXiv. 2023;abs/2310.11761. doi:10.48550/arXiv.2310.11761.

[15] Robinson J, Rytting C, Wingate D. Leveraging large language models for multiple choice question answering. ArXiv. 2022;abs/2210.12353. doi:10.48550/arXiv.2210.12353.

[16] Zhang Z, Lei L, Wu L, Sun R, Huang Y, Long C, Liu X, Lei X, Tang J, Huang M. SafetyBench: Evaluating the safety of large language models with multiple choice questions. ArXiv. 2023;abs/2309.07045. doi:10.48550/arXiv.2309.07045.

[17] Bitew SK, Deleu J, Develder C, Demeester T. Distractor generation for multiple-choice questions with predictive prompting and large language models. ArXiv. 2023;abs/2307.16338. doi:10.48550/arXiv.2307.16338.

[18] Ministry of Examination, Taiwan. “Exam Question and Answer Search System,” *Ministry of Examination*, Available at: https://wwwq.moex.gov.tw/exam/wFrmExamQandASearch.aspx?y=2012&e=101120, accessed on August 8, 2024.

[19] Nay JJ, Karamardian D, Lawsky S, Tao W, Bhat MM, Jain R, Lee AT, Choi JH, Kasai J. Large language models as tax attorneys: A case study in legal capabilities emergence. ArXiv. 2023;abs/2306.07075. doi:10.48550/arXiv.2306.07075.

[20] Phogat KS, Harsha C, Dasaratha S, Ramakrishna S, Puranam SA. Zero-Shot Question Answering over Financial Documents using Large Language Models. ArXiv. 2023;abs/2311.14722. doi:10.48550/arXiv.2311.14722.

[21] Kojima T, Gu S, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. ArXiv. 2022;abs/2205.11916. doi:10.48550/arXiv.2205.11916.

[22] Cheng D, Huang S, Bi J, Zhan YW, Liu J, Wang Y, Sun H, Wei F, Deng D, Zhang Q. UPRISE: Universal Prompt Retrieval for Improving Zero-Shot Evaluation. ArXiv. 2023;abs/2303.08518. doi:10.48550/arXiv.2303.08518.

[23] Ni Y, Jiang S, Wu X, Shen H, Zhou Y. Evaluating the robustness to instructions of large language models. ArXiv. 2023;abs/2308.14306. doi:10.48550/arXiv.2308.14306.

[24] Xu C, Sun Q, Zheng K, Geng X, Zhao P, Feng J, Tao C, Jiang D. WizardLM: Empowering large language models to follow complex instructions. ArXiv. 2023;abs/2304.12244. doi:10.48550/arXiv.2304.12244.

[25] Hugging Face. ”Open LLM Leaderboard MMLU,” *Hugging Face Blog*, Available at: https://github.com/huggingface/blog/blob/main/open-llm-leaderboard-mmlu.md, accessed on August 8, 2024.

[26] Liu Yu-Wei. ”LLM Model Evaluation,” *GitHub Repository*, Available at: https://github.com/LiuYuWei/llm_model_evaluation, accessed on August 8, 2024.

[27] MediaTek Research. “Breeze-7B-Base-v1_0,” *Hugging Face Model*, Available at: https://huggingface.co/MediaTek-Research/Breeze-7B-Base-v1_0, accessed on August 8, 2024.

[28] Yenting Lin. ”Llama-3-Taiwan-8B-Instruct,” *Hugging Face Model*, Available at: https://huggingface.co/yentinglin/Llama-3-Taiwan-8B-Instruct, accessed on August 8, 2024.

[29] Pei-Yuan Liu. ”MMLU Dataset,” *Kaggle Dataset*, Available at: https://www.kaggle.com/datasets/peiyuanliu2001/mmlu-dataset, accessed on August 8, 2024.

[30] iKala. ”TMMLUPlus Dataset,” *Hugging Face Dataset*, Available at: https://huggingface.co/datasets/ikala/tmmluplus, accessed on August 8, 2024.

Collision Care Guide based on Large Language Models
基于 LLM 的交通事故陳述記錄輔助助理

龔若齊 **Jo-Chi Kung** 張嘉惠 **Chia-Hui Chang**
黃懷萱 Huai-Hsuai Huang 簡國峻 Kuo-Chun Chien
國立中央大學資訊工程學系
Department of Computer Science & Information Engineering,
National Central University
z1a2x3s4c5d6v7f8b9g@gmail.com, chia@csie.ncu.edu.tw
chrbezz0487@gmail.com, qk0614@gmail.com

摘要

本研究提出了一個基於大型語言模型（LLM）的交通事故陳述記錄助理系統（CCG），旨在協助車禍當事人釐清事故經過，減輕警方與保險專員詢問的負擔。CCG 系統的核心設計包括三個主要模組：提問模組、資訊擷取模組和事故經過生成模組。提問模組利用預設的問題模板引導用户提供完整資訊；資訊擷取模組將用户回答準確擷取至預先定義的資訊儲存格式（TARF）中；事故經過生成模組則將結構化資料轉化爲連貫的事故敘述。這種模組化設計實現了高效、準確的事故資料收集和處理。

實驗結果顯示，CCG 系統在人工評估中 F1 分數達到 0.909，在 LLM 對話評估中準確性和完整性評分均達到 10 分中的 7 分以上。我們同時採用人工評估和 LLM 自動評估方法，驗證了系統在實際應用中的有效性。這些結果證明 CCG 系統具有良好的實用性和應用前景，能夠協助事故當事人精確記錄所提供的資訊，爲後續的法律程序和保險理賠提供可靠依據。

關鍵字：大型語言模型、對話系統、資訊擷取、交通事故

1 Introduction

根據台灣交通部路政及道安司的統計，臺灣交通事故數量逐年上升，從 2019 年發生 34 萬件事故，至 2023 年統計已突破 40 萬件，平均每天超過 1000 件事故發生[1]。當事人在發生交通事故後，通常會立即報案並聯繫保險公司，但警方與保險公司在處理時，首要任務是釐清事故經過，然而常見問題包括：當事人描述不完整或遺漏關鍵事實，導致責任歸屬難以確定；警力與保險公司人力有限，難以同時服務所有當事人，造成漫長等待；而大多數人在事故發生後往往不知如何應對，需仰賴有經驗者協助處理賠償與責任問題。在現今基於大型語言模型（LLM）的智能對話系統雖廣泛應用於教育與醫療領域 (Dam et al., 2024)，但針對交通事故領域的應用卻相對稀少。目前在交通事故領域的研究，多集中於事故預測與數據分析 (Behboudi et al., 2024)，針對當前事故對於事故當事人的記錄、蒐集信息等方面的研究也是較爲缺乏。

鑑於大型語言模型 (LLM) 在對話能力與資訊整合上的進步，我們選定由 OpenAI 所提出的 GPT 爲核心設計，因爲他在對話生成與訊息理解廣爲人知與表現出色。我們提出了一個基於 ChatGPT 的智能聊天系統 CollisionCare Guide (CCG)，設計了一個完整的問答流程，包含了三個不同的模組: 提問模組、資訊擷取模組和事故經過生成模組，並使用自定義的 JSON 表單模板，資訊儲存格式 (TARF)，以引導當事人逐步記錄成結構化的事故細節，最終也可以透過模組轉化爲連貫的敘述形式，以便警方或保險理賠員進行責任判斷與賠償計算。此系統研究目的是爲了減少多次重複問答的負擔，提升事故處理效率。爲驗證整體 CCG 的效能，我們進行了人工評估與 LLM 自動評估。在人工評估中，測試人員根據判決書模擬當事人進行描述回答，最終 F1 分數達到 0.9，顯示 CCG 能準確針對使用者回覆的資訊，擷取至正確的欄位中，值得注意的是，AI 代理人在相同的測試方法中也達到了 0.9 的 F1 分數，這表明 AI 代理人與眞人測試員有高度的相似度，這一結果證實了 AI 代理人可以有效地替代眞人進行更多對話上的模擬測試。在 LLM 自動評估中，AI 代理人與 CCG 進行互動對話，利用 LLM 自動評估結果顯示其完整性與精準性達至約 7/10 分，證明 CCG 能有效引導當事人並正確擷取所提供的資訊。

我們此論文的貢獻如下:

- 我們提出一個交通事故代理人 CCG，結合擷取與問答模組的交互合作，與可自定

[1] https://ba.npa.gov.tw/statis/webMain.aspx?k=defjsp

義的事件資訊模板，將記錄轉化成結構化的表達方式和交通事件的完整敘述，具有高度的應用彈性。

- 我們設計並實施了一套創新的評估方法，結合人工評估和 LLM 自動評估，全面驗證了 CCG 系統在實際應用中的有效性和準確性。

- 我們的研究設計一套完整問答流程，爲交通事故資訊蒐集和處理提供了一個範例，展示了 LLM 在特定領域應用的潛力，爲未來在其他法律案件中的應用奠定了基礎。

2 Related Work

專門處理車禍案件的市場規模相當龐大，尤其在交通事故頻繁的地區。根據上述官方資料，台灣每年發生數萬起交通事故，這些事故涉及的法律問題和賠償需求促使車禍律師的需求不斷增加。此外，隨著人們對法律權益的重視程度提高，越來越多的人選擇在發生車禍後尋求專業律師的協助，以確保自己的權益得到保障。

國際上專門處理車禍案件的律師事務所如 Alexander[2]等，通過結合地理定位和在線聊天機器人，提供律師匹配與法律諮詢的服務，如 Accident Consults[3]、FindLaw[4]和 Legal010[5]。這些平台目標在於提供專業見解並幫助受害者爭取賠償。除此之外，一些系統也運用了人工智能技術來提高法律服務的效率。舉例來説，LAW-U (Socatiyanurak et al., 2021) 採用了有限狀態機控制對話流程，爲性犯罪受害者提供精確法律建議，並能準確推薦相關最高法院判決。而 DoNotPay[6]作爲「第一位機器人律師」，讓使用者能夠處理法律糾紛如爭議罰單和隱私保護等問題。

針對交通事故分析的技術也逐漸發展，AccidentGPT (Wu et al., 2024) 爲一個多模態基礎模型，能處理多種輸入數據，包括音頻、影像、文本等，並自動生成多任務分析結果，從重建事故過程到生成責任歸屬報告皆涵蓋其中。此外，基於 BERT 的大型語言模型也被應用於交通事故的嚴重程度分類 (Grigorev et al., 2024)。這項研究使用了超過 75 萬份事故敘述，並達到 84.2% 的預測準確率，強調了 LLM 在處理大型事故數據集時的潛力，且降低了計算複雜性，提升了可擴展性。

隨著 LLM 技術的進步，研究發現這些模型可以替代人類進行自動化評估。根據 Chiang and yi Lee (2023) 的研究，LLM 在語言生成任務的評估結果與人類高度相關，顯示了其在自動化評估中的潛力。同樣地， Lin and Chen (2023) 提出了一種基於 LLM 的自動對話系統評估方法，該方法與人類評估結果顯示高度一致性，進一步證明了 LLM 在對話系統評估中的應用價值。

綜上所述，雖然有關交通事故領域的研究相當稀少，不過現有的研究展示了在法律方面與交通事故後續處理的解決方案，以及對於大型語言模型的評估方法。然而，這些系統普遍受限於特定的領域如法條處理或是事故分析。本研究提出的 CCG 系統，通過結合大型語言模型的對話能力和信息擷取技術，旨在解決這些問題，提供更高效、更準確的交通事故處理解決方案。

3 Collision Care Guide (CCG) Agent

一般而言，交通警察處理車禍事故的筆錄過程主要是透過提問以及當事人的回答來獲取相關資訊，會依序向當事人提問以下有關交通事故的問題：

- 事故發生的時間地點、使用的交通工具
- 事故前的起點、行駛道路、行進方向、交通號誌、標誌是否清楚、和事故經過
- 天候、路況、交通流量、障礙物等
- 碰撞位置、車損情形、人員受傷情形

我們參考警察處理交通事故[7]筆錄時的標準程序，定義事故重點結構模板（包含事件細節及事件細節解釋），詢問車禍當事人各項細節，確保不會遺漏任何重要資訊。

本研究中，我們使用 OpenAI 提供的 GPT-3.5-turbo 作爲 CCG Agent 的核心基礎，結合設計的 prompt 與事故處理模板格式，來了解交通事故經過。

根據 OpenAI 提供的建議中[8]，比起在一個指令下執行兩種不同任務，將單一複雜任務拆分成兩項子任務，更能使大型語言模型更專注

[2] https://shunnarah.com/practice-areas/car-accident-lawyer/

[3] https://www.accidentconsults.com/

[4] https://lawyers.findlaw.com/

[5] https://laws010.com/blog/car-accident/car-accident-lawyer/car-accident-lawyer-01

[6] https://donotpay.com/

[7] https://td.police.gov.taipei/News_Content.aspx?n=9CDDA66829FF2249&sms=5E5FFE038245884F&s=9D88B5BC9452512F

[8] https://platform.openai.com/docs/guides/prompt-engineering

於當前任務，不僅能降低錯誤率，更易於控制與修正。

如圖 1所示，我們將 CCG 系統中處理交通事故的問答過程，分爲提問與資訊擷取兩個模組，前者任務是負責確認是否有無缺漏的資訊進而提出相關問題，後者任務是負責將使用者回覆的資訊，擷取資訊並正確紀錄至相應的問題欄位中。透過一問一答的方式和當事人互動，將這些車禍事故的回覆資訊擷取至自定義的儲存模板中。最後，當模板中所有所需資訊都獲得完後，透過事故經過生成模組還原成事故經過。

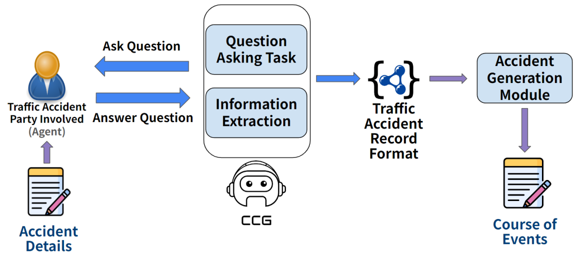


Figure 1: 多指令控制及事故經過生成流程

3.1 提問模組

在提問模組中，我們除了期望 CCG 能夠依照我們事先定義的資訊格式來詢問問題，也能夠依照當事人的狀況來適當地回覆。首先，CCG 要能夠根據當事人的回答，適切地以友善的語氣回應他們。如果當事人偏離了我們所要問的問題，CCG 會以專業且友善的態度引導當事人回到正題，正確地回答問題。如果當事人顯得緊張或不知所措，CCG 也能夠以安撫和鼓勵的話語幫助他們放鬆，並鼓勵他們提供所知的信息。當然若當事人正確地回覆了我們所提出的問題，CCG 也會適時地鼓勵他們繼續完成問答。

如同交通警察在筆錄時需要透過多輪對話了解事故經過一樣，CCG 的核心工作是提出問題並根據當事人的回覆獲取所需資訊。車禍事故記錄會以一個 JSON 格式來儲存，如表 2左欄交通資訊儲存格式 @TARF (Traffic Accident Record Format) 所示，這些交通資訊都是以警方會在車禍筆錄中所提問的常見問題，當然這些資訊都可以自行定義增減問題。另外，問題解釋格式 @QEF (Question Explaination Format) 中，則是對應著這一系列事故屬性鍵的名詞解釋。

在此任務中，我們會主動檢查資訊儲存格式 @TARF 中尚未提問的屬性鍵，然後將這些屬性鍵對應到問題解釋格式 @QEF 中的屬性解釋，與表 1中的第一欄 Prompt 組合後，輸入給大型語言模型，讓語言模型生成一個不僅能回應當事人的回覆，並且提出下一個問題資訊給予當事人作回答。

3.2 資訊擷取模組

當使用者回覆問題後，我們再透過資訊擷取提示，依靠著大型語言模型的統整能力，將使用者的答案填入相對應的資訊儲存格式中。如表1第二欄所示，資訊擷取提示除了擷取任務的提示詞之外，還包括使用者收到提問模組的上一則問題、使用者的回覆以及當下的資訊儲存格式 @TARF。上一則問題及使用者的回覆有助於大型語言模型更了解當下的事故擷取任務是針對哪一個資訊，從而可以正確擷取至對應到儲存格式中，並且輸出一個更新完整的資訊儲存 @TARF。

3.3 CCG 模型控制流程

CCG 控制流程如 Figure 1所示。首先，CCG 會詢問使用者事故發生時的情形概況，請使用者簡要描述案件狀況，包括發生日期、時間、地點和事發經過。接著，透過初始資訊提取提示，CCG 會將使用者提供的資訊輸出成預設的資訊儲存格式 @TARF，完成第一輪詢問及擷取。接著 CCG 會檢查資訊儲存格式 @TARF 中是否尚有空屬性值，加上屬性的定義後，CCG 透過提問模組提示產生一個問題給使用者。使用者回覆有關這個問題的資訊後，資訊擷取模組會根據使用者的回覆和資訊擷取提示擷取相關資訊，並將其對應到資訊儲存格式中正確的屬性鍵值上，然後更新格式，形成一次完整的問答。如果未能將交通事故記錄儲存格式中的所有資訊擷取出來，則會重複使用提問模組和屬性擷取模組，直到獲得所有所需的資訊。

3.4 事故經過生成模型

如 Figure 2所示，事故經過生成任務可以看成是交通事故資訊擷取問題的逆向工程，若能夠將這些原本在交通事故陳述上擷取下來的資訊用 Json 儲存，透過事實經過生成模組來還原回陳述形式的交通事故，就可以將使視爲正確完整的擷取。如 Table 1中第三欄所看到的，我們將 CCG 與當事人交互對話後，最終的資訊儲存格式 @TARF，將此 JSON 結果格式配合設計的提示詞，共同輸入給 GPT，就能得到敘述形式的交通事故陳述，整體流程如同 Figure 1後半部所示。最後不管是利用資訊儲存格式 @TARF 還是交通事故陳述，在後續不管是警方或是保險理賠人員，抑或是律師，都能透過兩種格式來因應不同場合需要，能夠

模組	**Prompts**
提問模組	作爲一名車禍事故陳述輔助專家，你的兩項主要任務是： 1. 根據 [上一個問題] 與 [當事人回答]，適當的回應當事人。 2. 根據 [下一個欄位]，提出下一個詢問的問題。 若當事人回答不相關，友善提醒他們專注於問題。對緊張的當事人，提供鼓勵。 每次回答後，根據 [問題解釋] 提出清晰的問題。 - [上一個問題]: {上一個問題} - [當事人回答]: {當事人回答} - [下一個欄位]: {下一個問題}的意思是{下一個問題解釋} - [專家的回覆]:
資訊擷取模組	你是資訊擷取機器人，請從 [問題回覆] 中擷取值填入 [Json 格式]。 - [Json 格式]:{'事故發生日期': '108 年 10 月 29 日','事故發生時間': '18 時左右', '事故發生地點': '', '我方駕駛交通工具': '', ...,'我方出發目的是什麼': ''} - [問題]:車禍發生時您駕駛的是什麼交通工具? - [問題回覆]:我當時駕駛的是一輛車牌號碼 000-0000 號的自用小客車。 - [Json]:
事故經過生成模組	你是一個車禍事故陳述記錄專家，根據 [Json 格式] 的事實，敘述車禍經過。 只描述 Json 中提供的事實，不包含其他資訊。 [Json]:{'事故發生日期': '108 年 10 月 29 日','事故發生時間': '18 時左右', '事故發生地點': '高雄市鳥松區松藝路與圓山路口', ...,'我方出發目的是什麼': '未知'}

Table 1: 各模組之提示詞 (藍色字體每輪對話會替換不同欄位資訊)

@TARF 資訊儲存格式	**@QEF** 問題解釋格式
"事故發生日期": "",	"事故發生日期": "具體日期",
"事故發生時間": "",	"事故發生時間": "具體時間",
"事故發生地點": "",	"事故發生地點": "具體地址",
"我方交通工具": "",	"我方交通工具": "交通工具種類",
"對方交通工具": "",	"對方交通工具": "交通工具種類",
"我方行駛道路": "",	"我方行駛道路": "行駛的道路",
"我方行進號誌": "",	"我方行進號誌": "號誌狀態",
"事發經過": "",	"事發經過": "詳細經過",
"當天天候": "",	"當天天候": "天氣情況",
"道路狀況": "",	"道路狀況": "道路狀況",
"我方行車速度": "",	"我方行車速度": "行駛時的車速",
"我方車損情形": "",	"我方車損情形": "車輛損壞情況",
"我方傷勢": "",	"我方傷勢": "傷勢情況",
"對方車損情形": "",	"對方車損情形": "車輛損壞情況",
"對方傷勢": "",	"對方傷勢": "傷勢情況",
"從哪裡出發": "",	"從哪裡出發": "出發地點",
"出發目的地": "",	"出發目的地": "目的地",
"出發目的爲何": "",	"出發目的爲何": "出發目的"

Table 2: 左邊爲資訊儲存格式 @TARF，右邊爲問題解釋格式 @QEF。

快速獲得當事人在此次車禍事故中的資訊，不必重複的詢問，浪費雙方的寶貴時間。

3.5 系統介面介紹

我們設計的 CCG 系統的介面預覽如圖 3 所示。此系統是通過結合多個模組，包括提問模組、資訊擷取模組以及事故經過生成模組，來實現與使用者的互動對話。本系統包含三個核心組件。當使用者發起對話時，系統首先會通過提問模組向使用者提出問題。這些問題根據預設的交通事故記錄模板設計，旨在收集詳細且準確的事故信息。提問模組能夠靈活應對使用者的各種回應，引導他們提供必要的細節。使用者回答問題後，資訊擷取模組會即時分析並擷取關鍵信息，將這些信息儲存在相應的資訊儲存格式中。這一模組利用大型語言模型的強大能力，確保信息的準確性和完整性。當所

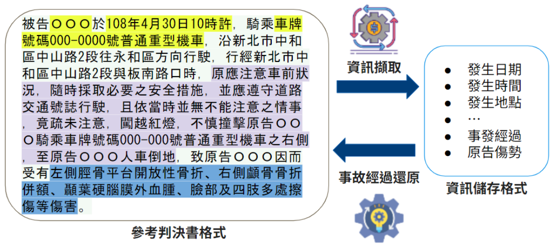


Figure 2: 交通事故資訊擷取 vs. 事故經過生成任務定義

有必要信息收集完成後，事故經過生成模組會根據儲存的資料生成完整的事故陳述。這一過程包括對信息的整合和重建，生成一份詳細的事故報告，供警方或保險理賠員使用。

這套模組化系統通過整合提問模組、資訊擷取模組和事故經過生成模組，提供了一個高效、準確的交通事故信息記錄與報告平台。這一系統不僅能夠提高事故處理的效率，還能顯著減少警方和保險理賠員的工作量，爲交通事故處理帶來革新性的改變。

4 Experiments

我們期望 CCG 在與當事人對話時，不僅可以使用人性化的語句詢問當事人，還能夠將當事人提供的車禍資訊準確地記錄到資訊儲存格式中。爲了達成這一目標，我們採用了兩項方式來評估：人工評估與 LLM 對話自動評估。在 CCG 中，最終目的也是需要眞實人類的使用，因此我們讓人類受測員參考判決書內容進行回覆，並以精確度 (Precision)、召回率 (Recall) 以及 F1 分數 (F1 Score) 爲指標，這

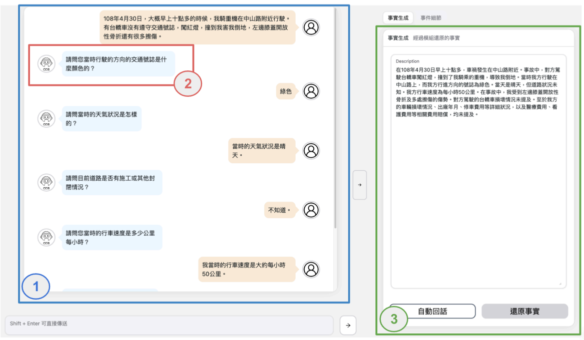


Figure 3: CCG 系統架設畫面預覽，(1) 爲對話內容，根據聊天內容藉由資訊擷取模組產生交通事故資訊儲存格式。(2) 的對話內容根據提問模組產生。(3) 爲事故經過生成模型。

些指標能全面反映 CCG 在資訊擷取的準確性與完整性。此外，我們也在人工評估中驗證 AI 當事人的表現，與眞實受測員的相似度，確保其能夠準確模擬眞實場景。

在 LLM 對話自動評估中，我們依照過去的研究如 LLM-Eval 自動評估的做法 (Lin and Chen, 2023)，利用大型語言模型資料分析的能力，來協助評估擷取與完整性這兩個指標任務。因爲我們本身 CCG 系統是採用 GPT，爲了避免潛在的偏見，在評估階段分別使用 GPT 和 Gemini 兩種 LLM 來進行獨立的綜合評估，以避免依賴單一 LLM 的評估結果來做出定論。我們希望這樣的綜合評估能夠更全面地了解 CCG 在與 AI 代理人對話中的表現，以及對車禍資訊的準確記錄能力，以獲得更具代表性和可信度的評估結論。

4.1 Human Evaluation

在本次實驗中，我們針對人類代理人評估進行設計，旨在評估 CCG 系統在眞實場景中的表現。因爲在現實我們無法取得眞實車禍的資訊，因此讓受測員扮演了參考判決書中車禍場景的原告，並且依照原告的主述，誠實不虛假的回覆每個 CCG 的問題，就像在眞實與警察做筆錄時需要誠實的回答所需的資訊。此實驗方法考慮了 CCG 是否正確擷取了當事人提供的信息，但在參考判決書中的原告敘述內容通常只會針對事發經過與傷害損傷賠償爲重點做論述，大多資訊如天氣、道路狀況與行車目的等內容並未提及，在目前我們自定義的 18 個欄位中平均約只有 8 至 10 個欄位有提及。因此我們在評估計算時，會以每筆判決書中提及的欄位內容爲主，其餘未談及的欄位將不列入計算。

實驗設計流程包括以下步驟：

1. 資料準備：首先，我們使用 GPT-4o 對判決書共 25 篇進行資訊擷取至 @TARF 中，並由人工檢查確認擷取結果的正確性，這些人工修正後的結果作爲參考。這些正確擷取的欄位數（去除未提及的欄位）記錄爲 a_i。

2. 受測員問答：受測員會隨機抽取判決書內容，模擬當事人進行回答，並提供相關的車禍信息，CCG 系統根據受測員的回答進行擷取。CCG 詢問的欄位數（去除 GPT 未擷取的欄位）記錄爲 q_i。

3. 數據記錄：最終記錄每次受測員與 CCG 對話後，正確擷取至資訊儲存格式 @TARF 的欄位數（去除 GPT 未擷取的欄位）記錄爲 e_i。

4. 計算指標：根據下述公式計算精確度（Precision）、召回率（Recall）和 F1 分數（F1 Score）。

- 精確度（**Precision**）：

$$P(i) = \frac{e_i}{q_i}$$

  精確度表示在 CCG 擷取出的所有欄位中，正確的比例是多少。

- 召回率（**Recall**）：

$$R(i) = \frac{e_i}{a_i}$$

  召回率表示所有應擷取的欄位中，CCG 成功擷取的比例是多少。

- **F1** 分數（**F1 Score**）：

$$F(i) = \frac{2 \cdot P(i) \cdot R(i)}{P(i) + R(i)}$$

  F1 分數是精確度和召回率的調和平均數，提供了一個整體的性能評估。

4.1.1 AI Agent (GPT Agent)

爲了更全面地評估 CCG 的表現，除了眞人受測員，我們同時使用 GPT-3.5-turbo 作爲 AI 代理人進行模擬測試。我們設定 AI 代理人參考與人類代理人相同的判決書內容，並依據經過精心設計，多次迭代的提示詞如 Table4下，確保 AI 代理人會依據判決書的內容，如實的描述判決書上的內容來進行測試。AI 代理人進行的結果，將與眞人測試結果進行比較，評估 AI 代理人是否能與人類受測員結果相近，來進行大規模的對話測試，以更好的評估 CCG 的表現。

4.1.2 結果

我們共蒐集了 25 筆對話，分別使用人類代理人和 AI 代理人進行測試。平均每筆對話的對話輪次爲 12.68 次（人類代理人）和 10.92 次（AI 代理人），對話輪次意思也就是使用者回答一句，加上 CCG 回覆一句，兩句話爲一輪次。結果顯示，CCG 系統在人類代理人測試中的精確度爲 0.923，召回率爲 0.897，F1 分數爲 0.909；而在 GPT 代理測試中的精確度爲 0.919，召回率爲 0.900，F1 分數爲 0.908。從結果中可以看出，CCG 在處理交通事故相關信息時，無論是面對人類代理人還是 AI 代理人，在 F1 分數都超過了 0.9，這代表著只要在使用者能夠正確的敘述資訊時，CCG 就能夠有效地擷取大部分必要的欄位，準確並且完整的紀錄在正確的欄位上。這使得我們有信心的可以說明，CCG 系統在眞實場景中能夠實際應用的潛力，特別是在處理交通事故相關訊息時的特定領域。

AI 代理人和人類代理人在整體 F1 分數上非常接近，但我們觀察到在某些特定類型的問題上存在細微差異。例如，AI 代理人在回答關於精確時間和地點的問題時，只要能在判決書上呈現的都表現優異，而人類代理人在描述主觀感受時更爲自然，這些差異提醒我們在解釋結果時需要考慮 AI 代理人的局限性。雖然 AI 代理人足以驗證能進行大規模、低成本的對話模擬測試，但我們也認識到它可能無法完全捕捉人類行爲的複雜性和不可預測性。因此，我們將後續 AI 代理人在 LLM 自動評估視爲人類評估的延伸補充，而非替代。

	平均輪次	P	R	F1
人類 Agent	12.7	0.923	0.897	0.909
GPT Agent	10.9	0.919	0.9	0.908

Table 3: CCG 系統在人類代理人和 AI 代理人評估中的表現。P 代表 Precision、R 代表 Recall、F1 代表 F1 Score。

4.2 LLM Evaluation

在前面提到了 AI 代理人與人類代理人的相似性，因此爲了評估 CCG 系統，我們需要使用大量的對話資料。我們設定 AI 代理人如上述一樣，設定相同的提示詞與規則 Table4，與 CCG 進行對話互動，以測試 CCG 的擷取模組是否能夠精確地擷取 AI 代理人的回應並將其記錄到正確的欄位中，評估整體對話的完整性和準確性。我們使共使用 587 篇的判決書隨機選擇，使用其中的車禍事實陳述作爲 AI 代理人的輸入，將 AI 代理人模擬成判決書中的原告，讓其與 CCG 進行交互問答，講述車禍相關的資訊。在評估方面，我們設定了兩種評估指標，擷取評估與整體評估，前者專注於 CCG 在單句話的擷取表現，後者專注於最終 @TARF 格式中整體的擷取結果。

4.2.1 擷取評估

在擷取評估中，我們對每一次 CCG 與 AI 代理人的對話進行評估，專注於 AI 代理人根據判決書中的車禍事實所做的回答，以及 CCG 提出的問題是否準確地被擷取並記錄到相應的 @TARF 的空欄中。我們使用 GPT-3.5-turbo 與 Gemini-1.0-pro 兩個 LLM 做爲評估模型，每次對話都會做一次準確性評分，評分範圍爲 1 至 10 分，並取每次對話的平均值作爲對話準確性的評估指標。而擷取評估中的 prompt 如 Table4，並在每次對話中，CCG 的問題、AI 代理人的回答、以及擷取後的 @TARF，一同加入提示詞作爲輸入。爲了進行評估，我們共收集了 587 筆對話，而對話的平均輪次爲 10.8 次，如圖4中 (b) 部分所示。我們將所有對話的準確性評估平均分繪製成分布圖如圖4中 (a) 所示。根據我們的評估，GPT-3.5-turbo 和 Gemini 的平均分別爲 7.93 分和 8.16 分。總體而言，平均準確性達到 8 分，可以表示 CCG 在單句擷取當事人所提供的資訊，大致上可以準確地被擷取到正確的欄位中。値得注意的是，因爲參考判決書中部分資訊並未提供，例如天氣、道路狀況或交通號誌等等，所以 AI 代理人若因爲部分資訊未提供而無法回答的情形，我們的評分將其給予 5 分，即使是這樣的評分標準，大多數情況下，系統能夠精確地擷取所需資訊。

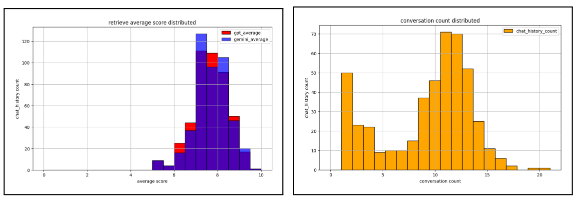


Figure 4: (a) 擷取評估準確性平均分數分布 (b) 對話輪次分布

4.2.2 整體評估

在整體評估中，我們的目標是確保最終的資訊儲存格式 @TARF 盡可能地與最初提供給 AI 代理人的判決書事故描述相符。爲了評估這一目標，我們對完整性和準確性分別給予了 1 至 10 分的評分。其中，完整性評分評估了是否將判決書事故描述中的所有資訊都完整地擷取至最終 @TARF 格式中，而準確性評分則評估了擷取到的資訊是否與判決書事故描述中所描述的一致且正確。我們一樣與上段擷取評

Agent 與評估	Prompts
AI 代理人	扮演 [車禍事故] 中的原告，簡答警察問題，未提及者回答' 不記得' 或' 忘記了'。 [車禍事故]:109 年 9 月 27 日 18:56，被告駕駛車牌 000-0000 自用車， 在新北市 0 路 0 段與華鋒三街口撞損原告車牌 000-000 重型機車，致修理費 14,200 元。 [警察的問題]:描述車禍當時的情況（時間、地點、經過等）
擷取評估	你是一位事故資訊分析師，依照以下內容確認使用者回答是否正確擷取到 Json 欄位。 ** 評分準則:** 準確性（1-10 分），資訊錯置扣分，回答不知情者 5 分。 ** 對話內容:** **CCG Agent 問題 **: CCG 提問 ** 使用者回答 **: User 回答 **Json**: @TARF ** 輸出格式:** 擷取評分: 整數分數 解釋說明: 指出正確與錯誤擷取，解釋分數。
整體評估	依照事發經過敘述與 JSON 資料，分析事故敘述與 Json 資料的匹配度， 基於以下標準給出完整性和準確性的整數分數（0 到 10 分）及詳細的評分理由。 ** 輸出格式:** ** 標準:** 完整性分數（根據敘述捕捉關鍵點），準確性分數（擷取內容是否匹配）。 ** 原因解釋:** 說明完整性和準確性評分理由，僅針對敘述提及項目。 ** 輸入:** ** 事發經過敘述 **：參考判決書事發經過 **JSON 資料 **：最終 @TARF 格式

Table 4: AI 代理人提示詞以及擷取評估、整體評估的提示詞。藍色字體在每輪會替換成不同的欄位資訊。

	對話總量	平均輪次	GPT	Gemini
值	587	10.8	7.93/10	8.16/10

Table 5: 對話評估結果

估相同，共收集了 587 筆對話，對話的平均輪次爲 10.8 次，使用了 GPT-4 和 Gemini-1.0-pro 進行綜合評估和比較。根據我們的評估結果，在準確性的部分，分布如圖5中所示，GPT 的準確性評分平均爲 7.41，Gemini 爲 8.31；在完整性的部分，GPT 的完整性評分平均爲 6.76，Gemini 爲 7.59，可以看到大部分的分數都落在 7 至 9 分，Gemini 給予較 GPT 高一點的分數，這表明 CCG 系統能夠大致上將 AI 代理人所敘述的資訊完整且準確地記錄在資訊儲存格式 @TARF 中。

模型	平均準確性分數	平均完整性分數
GPT-4	7.41	6.76
Gemini	8.31	7.59

Table 6: 準確性和完整性評估結果

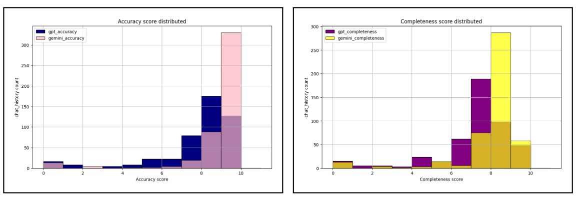


Figure 5: (a) 整體準確性評估分布 (b) 整體完整性評估分布

5 Limitation

在本研究中，我們設計並評估了 CCG 系統，但仍然存在一些限制需要考慮。

- 隱私問題：我們著重於保護當事人的隱私，並不會蒐集使用者的個人資料，若要在某些情況下需要蒐集，必需要進一步的數據隱私保護措施並告知使用者，以確保所有收集到的信息都得到妥善處理。
- 無法進行法律判決：CCG 系統的設計重點在於紀錄當下的車禍事件。在法律判決或賠償金額的決定上，CCG 系統無法替代專業的法律判斷，需要綜合雙方的證詞、影像、客觀錄音以及專家意見才能進行評估。
- 欄位信息不全：我們的實驗主要依賴於模擬環境和預定的判決書資料，在這之中部分欄位如天氣、道路狀況、行車目的等並無提及，這可能無法完全反映眞實世界中的多變情況。
- AI 代理人侷限性: 雖然 AI 代理人在我們的實驗中表現良好，但它們可能無法完全模擬眞實人類在壓力下的反應或不一致的回答，這可能導致在實際應用場景中出現一些未預期的情況。
- 實驗環境局限性：我們當前的實驗主要集中在交通事故場景，且基於中文的特定語言與文化上，在其他法律或非法律領域，

與不同語言和文化的應用效果尚未驗證，這是未來研究可以嘗試的方向。

這些限制說明了目前我們研究中的挑戰，並且提供未來改進的方向，爲後續的研究提供了參考。

6 Conclusion

CCG 系統旨在幫助當事人在車禍事件後，能夠盡速的協助當事人與警方紀錄相關的事件細節，減少重複繁瑣的提問過程，在日後車禍事故判斷或是賠償金額的計算前不必再次詢問。此外，透過我們設計的問答模式與模板設計，都可以根據實際需求進行調整和擴展，在未來可應用至更多法律問答機器人或非法律領域，例如民事、刑事案件的處理，或企業收集用户資訊等應用。

在實驗中，CCG 能夠依照所定義設計好所需的事件細節，準確的紀錄使用者所回答的資訊，進行了人工評估與 LLM 自動評估。在眞人測試中，眞人代理人和 AI 代理人測試與 CCG 的模擬對話中，CCG 系統在資訊擷取能力上表現優良，精確度、召回率和 F1 分數均落在約 0.9 上下。在 LLM 自動評估中，通過 GPT-4 與 Gemini-1.0-pro 的綜合評估下，CCG 在擷取評估與整體評估中也都達到約 7 分的標準，這顯示當使用者誠實的回覆問題下，我們不僅證明 CCG 能有效處理交通事故紀錄，即便在部分如天氣、道路狀況等資訊缺乏的情況下也拿精準將訊息擷取至正確的欄位中，也確保了評估的客觀性與可靠性。然而，我們認識到 AI 代理人能使我們進行大規模、低成本的測試，但 AI 代理人無法完全替代眞人測試，兩者應該相輔相成。

總結來說，CCG 系統展示了在交通事故處理中的潛力，但也存在改進空間。未來的工作將集中在提高 CCG 系統的整體程度，更能夠友善並有效地協助當事人，處理更多樣化的事故場景。我們也將致力於擴大眞實世界的測試範圍，適應更多樣化的事故場景，並探索 CCG 在其他法律和非法律領域的應用可能性，如民事糾紛、醫療諮詢等。本研究不僅爲交通事故信息收集提供了一個創新的解決方案，也爲人工智能在法律和公共服務領域的應用開闢了新的可能性。

References

Noushin Behboudi, Sobhan Moosavi, and Rajiv Ramnath. 2024. Recent advances in traffic accident analysis and prediction: A comprehensive review of machine learning techniques.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations?

Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A complete survey on llm-based ai chatbots.

Artur Grigorev, Khaled Saleh, Yuming Ou, and Adriana-Simona Mihaita. 2024. Enhancing traffic incident management with large language models: A hybrid machine learning approach for severity classification.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Vorada Socatiyanurak, Nittayapa Klangpornkun, Adirek Munthuli, Phongphan Phienphanich, Lalin Kovudhikulrungsri, Nantawat Saksakulkunakorn, Phonkanok Chairaungsri, and Charturong Tantibundhit. 2021. Law-u: Legal guidance through artificial intelligence chatbot for sexual violence victims and survivors. *IEEE Access*, 9:131440–131461.

Kebin Wu, Wenbin Li, and Xiaofei Xiao. 2024. Accidentgpt: Large multi-modal foundation model for traffic accident analysis.

Reformulating Programming Assignments: Balancing Correctness and LLM Resistance

Hsin-Chih Ho, Sin-Syuan Wu Yao-Chung Fan
Department of Computer Science and Engineering,
National Chung Hsing University, Taiwan

Abstract

This research addresses the challenge of creating effective programming assignments in the era of large language models (LLMs). Teachers can now use LLMs like ChatGPT to generate assignments, but students may also rely on these models to solve them. To mitigate this, we propose an Automated Algorithmic Question Generation (AAQG) method that augments dataset information to create scenario-based questions that challenge LLMs.

We introduce two new metrics: Topic Integrity Score (TIS) and Knowledge Similarity (KS). These evaluate the differences and core knowledge between generated and original problems. Our experiments using the LeetCode-sourced TACO dataset show improvements in TIS by up to 0.167 points and KS by up to 0.084 points over the baseline. Furthermore, Pass@k evaluation demonstrated our method's effectiveness in disrupting LLM performance, with GPT-3.5-turbo-0125 experiencing a 42-point drop in Pass@1 and CodeLlama-13b-Instruct-hf showing a 3.4-point drop in Pass@1 and a 7.21-point drop in Pass@5.

Keywords: LLM, Programming Education, Question Generation

1 Introduction

In computer programming education, teachers typically need to design programming assignments (as shown in table 1) to assist students in their learning. To improve the efficiency of assignment creation, teachers often modify existing past exam questions. However, with the advancement of large language models (such as ChatGPT and similar tools), teachers can combine past exam questions with new ideas and input them into the model to help generate new assignments. Nevertheless, when using large language models to generate programming assignments, teachers may face the following challenges.

- Correctness: The generated problem may lack a solution due to missing information or incoherent logic in the prompt. Additionally, the generated result might deviate from the original topic. For example, if the input is a sorting-related problem, the language model's output might reduce the problem to just a general array concept, straying away from the intended subject.
- Difficulty: During the learning process, it is difficult for teachers to prevent students from using large language models to solve their assignments. Therefore, teachers also hope that the problems they generate will be challenging for the language model to solve easily.

With the rise of large language models, their remarkable performance in various natural language processing tasks has garnered significant attention. In recent years, researchers (Austin et al., 2021; Hendrycks et al., 2021; Li et al., 2022; Nijkamp et al., 2022; Li et al., 2023a; Roziere et al., 2023; Li et al., 2023b) have begun focusing on the capabilities of large language models in code generation. Since the development of these models, their ability to generate code has seen notable improvement, particularly in solving algorithmic problems, where they have demonstrated great potential.

However, previous researchers have primarily focused on generating high-quality code and have yet to explore the generation of pro-

Question	Count the number of prime numbers less than a non-negative number, n. Example: Input: 10 Output: 4 Explanation: There are 4 prime numbers less than 10, they are 2, 3, 5, 7.
Skill types	Data structures, Range queries

Table 1: Example of a Programming Assignment: In the question description, in addition to the question itself, a set of test data is also included. In this example, the required problem-solving technique involves data structures and range queries.

gramming problems. Traditionally, programming problems rely on domain experts, such as teachers, to manually create them, which is both time-consuming and labor-intensive. Therefore, we aim to develop an automated algorithmic problem generation system to produce high-quality problems.

In practical terms, with the development of large language models, teachers are facing a new challenge: students are increasingly using tools like ChatGPT to complete algorithm assignments and write code. Therefore, we aim to generate problems that can effectively interfere with language models, making it difficult for them to provide easy solutions.

This study proposes an Automated Algorithmic Question Generation (AAQG) method. AAQG combines existing problems with their core problem-solving concepts and uses large language models to generate new algorithmic problems that share the same problem-solving core as the input problem. To evaluate the quality of the generated problems, we designed two assessment metrics specifically for algorithmic problems: Topic Integrity Score (TIS) and Knowledge Similarity (KS). The TIS metric assesses the differences between the generated and original problems, while the KS metric evaluates the similarity in core concepts between the generated and original problems.

In our experiments, we used a subset of the TACO dataset sourced from LeetCode and evaluated problem quality using our proposed TIS and KS metrics. We employed the Pass@k metric from (Chen et al., 2021) to observe whether large language models showed signs of interference, resulting in reduced problem-solving capabilities. The results indicate that our method improved the TIS score by up to 0.167 points and the KS score by up to 0.084 points. Furthermore, our method effectively reduced the language models' performance on Pass@k. When using CodeLlama-13b-Instruct-hf as the code generation model, Pass@1 dropped by up to 3.4 points and Pass@5 decreased by up to 7.21 points. When using GPT-3.5-turbo-0125, the reduction was even more significant, with Pass@1 dropping by up to 42 points.

2 Related Work

In this section, we primarily discuss the details of how previous researchers have utilized large language models for code generation.

Code Generation Model In the past, there have been numerous studies on code generation, such as (Austin et al., 2021; Hendrycks et al., 2021; Li et al., 2022; Nijkamp et al., 2022; Li et al., 2023a; Roziere et al., 2023), which focus on training new models for the task of code generation. Among these, (Roziere et al., 2023) introduced Code Llama, an open foundational model specifically designed for code generation.

Code Llama is notable for its robust multi-turn code generation capabilities, which make it particularly effective at handling complex algorithmic tasks. Built on the latest deep learning technologies, Code Llama can generate high-quality code and support a wide range of code generation tasks, from simple to complex. Its open nature provides broad applicability in the research community and facilitates easy integration into various programming-related applications.

In (Li et al., 2022), the focus is on generating code for competitive programming problems. The paper introduces AlphaCode, a system designed to generate high-quality competition-level code. AlphaCode aims to participate in programming contests and produce code that meets competition standards. The authors present the model architecture, training process, and performance of AlphaCode in programming contests. Experimental results demonstrate that AlphaCode excels in generating code of competitive programming quality.

Code Generation Dataset In (Li et al., 2023b), a novel dataset called TACO was introduced for the code generation task. This dataset includes problems sourced from major online programming platforms such as LeetCode and Codewars, and it also features difficulty levels for the problems. The paper evaluates various large language models on this dataset, with GPT-4 showing the best performance, exceeding other large language models by up to 22.5 points.

Since the TACO dataset provides comprehensive information on algorithmic problem generation, we utilized it for our problem generation experiments. We also referenced the comparison subjects from the research to guide our experiments. For code generation models in our experiments, we used Code Llama and GPT.

3 Method

Algorithmic problem generation is an emerging research field, and as a result, there has been a lack of robust evaluation methods to assess the quality of generated problems. This section will introduce our proposed problem generation method and the design of automated metrics to evaluate different aspects of problem quality.

Therefore, in this section, we will cover: (1) how our method, Automated Algorithmic Question Generation (AAQG), automatically generates algorithmic problems, and (2) how we validate the quality of the generated problems.

3.1 AAQG

In this paper, we primarily use existing problems and their problem-solving techniques as inputs to automatically generate new problems related to the input through our automated process.

Figure 1 presents an overview of the entire AAQG process. We utilize a large language model (LLM) to accomplish our generation task. AAQG is divided into three main components: Query Expansion, Retrieve, and Question Generation.

3.1.1 Query Expansion

In this section, we use large language models to expand the input problems and their problem-solving techniques through prompting. This process is mainly divided into Question Expansion and Skill Type Expansion, as illustrated in Figure 2.

Question Expansion Since the ultimate goal is to generate new problems related to the input problem, we first input the problem into the LLM for data expansion. The LLM generates application scenarios related to the problem, including a "title" and "description." This approach aims to provide context for the subsequent problem generation, making the generated problems richer and more contextualized.

Skill Type Expansion Since the problem-solving techniques in the dataset are often broad topics, such as data structures or dynamic programming, we aim to make the generated problems more focused. To achieve this, we input the problem and its associated techniques into the LLM. The LLM then supplements this information by identifying additional problem-solving techniques involved or specifying subtopics within the given technique, and describes the relevant problem-solving knowledge.

3.1.2 Retrieve

There have many past studies discuss the issue of hallucinations in large language models. To ensure that the generated problems are more realistic, we use the "title" from the application scenarios generated by the LLM as input

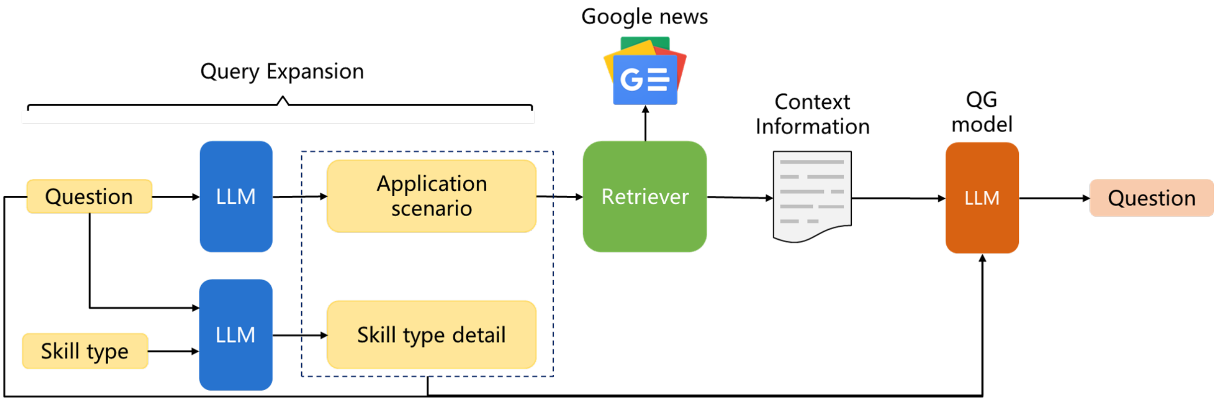


Figure 1: Overview of Methodology

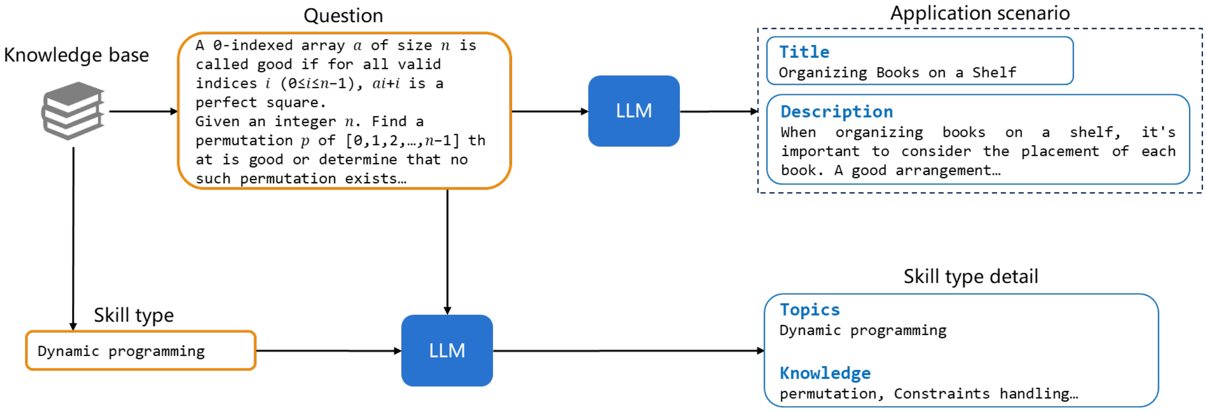


Figure 2: Query expansion

to retrieve the most relevant article. The goal is to make the content more grounded in reality. We use Google News as the retriever database to obtain a news article related to the problem, which serves as a reference for the subsequent problem generation.

3.1.3 Question Generation

In this section, we use the problems from the dataset, previously generated application scenarios, skill type detail, and a context information as inputs. We then employ the LLM to generate problems related to these inputs and create test data for the problems, as illustrated in Figure 4.

In the question generation phase, we use a two-stage generation approach. The first stage focuses on generating the questions, while the second stage is responsible for generating the test data for the questions. This two-stage approach is employed because generating both questions and test data simultaneously might distract the model from the current task, potentially leading to incoherent test data. Therefore, we choose to separate the generation of questions and test data.

Since subsequent validation will be conducted using unit tests, the LLM generates test data focusing on the following aspects: baseline testing, boundary testing, and random testing. As the generated problems do not currently account for execution time or memory requirements, stress testing is not specifically included in the test data generation. Ultimately, this process generates 3 to 10 test cases, with at least one test case for each of the baseline, boundary, and random tests.

3.2 Token Score For Question Generation

This section introduces the two evaluation metrics, Topic Integrity Score and Knowledge Similarity. These metrics assess the similarity between the generated problem and the original input problem in terms of their core problem-solving concepts.

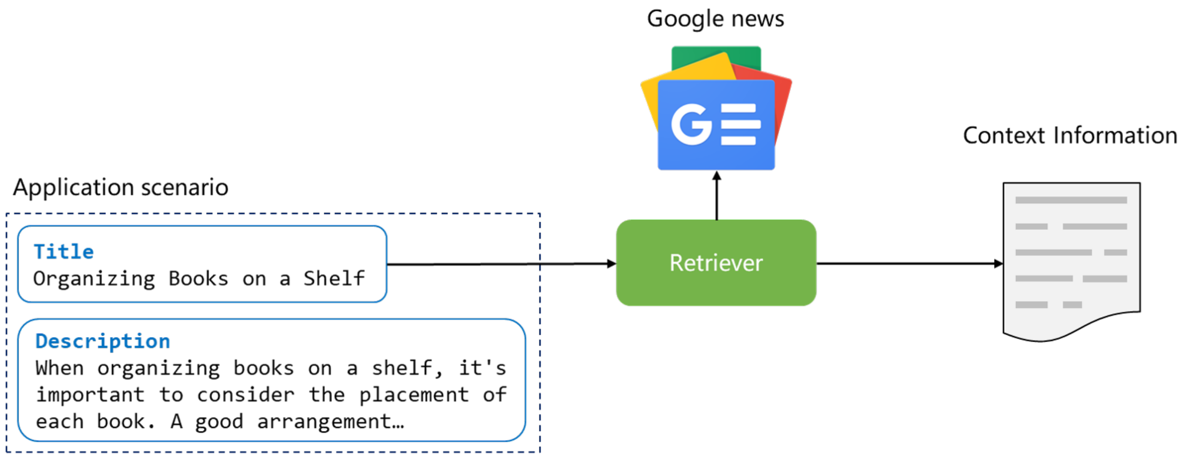


Figure 3: Retrieve

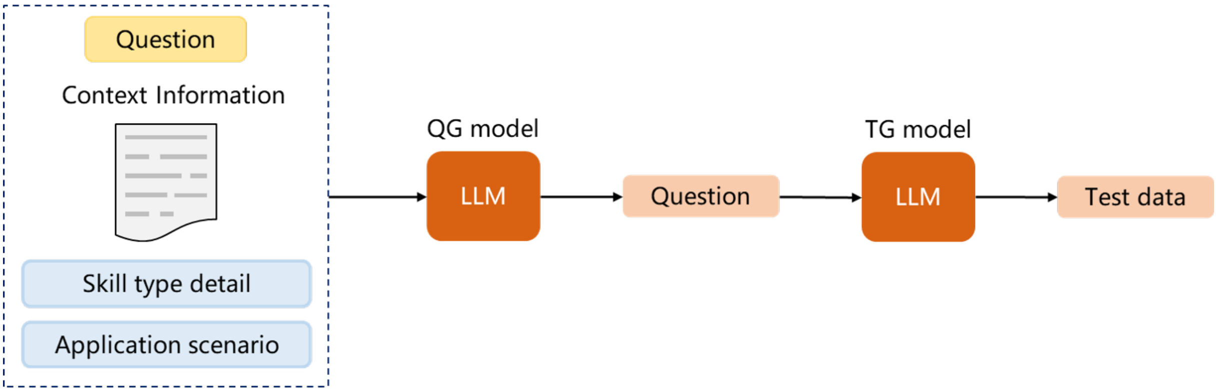


Figure 4: Query generation

3.2.1 Topic Integrity Score

Topic Integrity Score (TIS) is a comprehensive evaluation metric used to calculate the lexical and intrinsic differences between the generated problem and the original problem. Its goal is to ensure that the generated problem retains core similarity to the original problem while maximizing differences in wording, as shown in Formula 1.If the generated problem is identical to the original problem, the TIS score will be 0. This is because we aim for significant lexical differences between the generated problem and the original problem, while still maintaining similarity in core content.The TIS score ranges from 0 to 1.

$$TIS = \begin{cases} 0 & \text{if } S^Q = 1 \\ \frac{1-S^Q+D^Q}{2} & \text{otherwise} \end{cases} \tag{1}$$

In the formula 2, q_o represents the original input problem, while q_g denotes the generated problem. The score D^Q reflects the similarity of the intrinsic core between the generated problem and the original problem, specifically the dense score of the two problems, calculated using cosine similarity.

$$D^Q = \frac{\vec{q_o} \cdot \vec{q_g}}{|\vec{q_o}||\vec{q_g}|} \tag{2}$$

In the formula 3, S^Q represents the lexical similarity between the original input problem and the generated problem, specifically the sparse score of the two problems. This is based on the ROUGE-L score from (Lin, 2004), with Formulas 4 and 5 corresponding to the recall and precision scores, respectively.

$$S^Q = \frac{(1+\beta)^2 R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \tag{3}$$

$$R_{LCS} = \frac{LCS(q_o, q_g)}{len(q_o)} \tag{4}$$

$$P_{LCS} = \frac{LCS(q_o, q_g)}{len(q_g)} \tag{5}$$

3.2.2 Knowledge Similarity

Knowledge Similarity (KS) is an evaluation metric designed to calculate the similarity of problem-solving knowledge between the generated problem and the original problem, as shown in Formula 6.The KS score ranges from 0 to 1.

$$KS = \frac{1 + w \times S^K + D^K}{2 + w} \tag{6}$$

In Formula 7, k_o refers to the detailed problem-solving knowledge obtained from the skill type expansion of the original problem, while k_g represents the problem-solving knowledge details generated directly by the LLM using the same prompt as in the skill type expansion. The D^K score aims to measure the internal similarity of the problem-solving knowledge between the two, specifically the dense score of the knowledge, calculated using cosine similarity.

$$D^K = \frac{\vec{k_o} \cdot \vec{k_g}}{|\vec{k_o}||\vec{k_g}|} \tag{7}$$

Given the extensive specialized knowledge involved in the algorithm domain, we also consider the sparse score S^K between k_o and k_g, as shown in Formula 8. In the formula , w represents the proportion of the sparse score to be considered, with w ranging between 0 and 1. Since the results of each generation may vary, we place more emphasis on the dense score.

$$S^K = \frac{(1+\beta)^2 R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \tag{8}$$

$$R_{LCS} = \frac{LCS(k_o, k_g)}{len(k_o)} \tag{9}$$

$$P_{LCS} = \frac{LCS(k_o, k_g)}{len(k_g)} \tag{10}$$

4 Experiment

In this section, we will introduce the dataset we used, the experimental setup details, the evaluation metrics, and the experimental results.

4.1 Dataset

Our experiments primarily utilized data from the TACO dataset, specifically focusing on problems sourced from LeetCode. The TACO dataset comprises problems from various platforms, each with its own unique difficulty definitions. To ensure the clarity and relevance of our experimental results, we chose to use data exclusively from LeetCode, categorized into three difficulty levels: Easy, Medium, and Medium_Hard, totaling 646 records. The detailed distribution is shown in table 2, which outlines the 646 problems divided into Easy, Medium, and Medium_Hard difficulty levels.

LeetCode	Easy	Medium	Medium_Hard
# of Questions	124	276	246

Table 2: The distribution of data from LeetCode within the TACO dataset, where the difficulty levels are limited to Easy, Medium, and Medium_Hard.

4.2 Implementation Details

In this section, we will provide a comprehensive description of the language models used and the code generation models employed for validation.

4.2.1 The Employed LLMs

- AAQG Model we use the GPT-3.5-turbo-0125 model developed by the OpenAI team. We access the model via API, set the temperature to 0.5, and utilize the function calling approach during operation.

- Code Generation Model we tested with the GPT-3.5-turbo-0125 model developed by OpenAI and the CodeLlama-13b-Instruct-hf model provided by Meta. The CodeLlama-13b-Instruct-hf model was downloaded and used through the Hugging Face platform.

- GPT-3.5-turbo-0125: We use the default temperature setting to generate code and unit test snippets.

- CodeLlama-13b-Instruct-hf: We set the temperature to 0.9 and top_p to 0.95 for generating code and unit test snippets.

4.3 Automatic Metric

We use our proposed TIS and KS metrics, as well as the pass@k metric introduced by (Chen et al., 2021), as evaluation indicators.

- TIS: This evaluation metric is used to assess the diversity of the generated questions in terms of format and their consistency in content.

- KS: This evaluation metric is used to assess the similarity in problem-solving knowledge between the generated questions and the original problems.

- Pass@k: This metric calculates the probability of the LLM correctly answering the questions.

4.3.1 Embeding Model

In the scoring metrics, TIS and KS, the score D involves converting sentences into embeddings and calculating their similarity. For the embedding conversion, we use the BGE-M3 model proposed by (Multi-Granularity), which demonstrates excellent performance in similarity scoring.

4.3.2 Baselines

In the experiments on question generation, there are two main parts.

- Differences Between LLM-Generated Questions and Original Questions: Since there is no existing research focused on problem generation in the algorithm domain, we use a direct approach where problems are input into the LLM, and problems are generated through prompting as our baseline.

- Impact on Language Model Answering Capabilities: In this experiment, we use problems from the TACO dataset as input for the language model to generate code, which serves as our baseline.

4.4 Experimental results

In this section, we primarily discuss two experiments: 1. The experiment on the differences between the LLM-Generated questions and the original questions. 2. The experiment on the impact of generated questions on the language model's answering capabilities.

4.4.1 Differences Between LLM-Generated Problems and Original Problems

In table 3, we present experiments conducted across three difficulty levels: Easy Medium and Medium_Hard.

The experimental data indicates that when problems are directly generated by the LLM, the TIS score is consistently 0. According to Formula 1, this implies that the sparse score of the generated problems is identical to that of the input problems, meaning that the generated problems are identical to the original ones. We speculate that this phenomenon may be due to the dominant influence of the input problems in the prompt, which causes the model to generate problems that are identical to the input ones.

We also compared the AAQG method with different configurations of AAQG, including settings without Skill Type Expansion and Retrieve. The experimental data shows that, although the TIS score for AAQG is 0.002 points lower than the configuration without the Retrieve component in the Medium difficulty level, the KS score is higher by 0.068 points. Thus, the AAQG method demonstrates the best overall performance.

The experiments also show that removing the Skill Type Expansion step leads to a significant decrease in both TIS and KS scores, with the KS score dropping by as much as 0.073 points. Therefore, Skill Type Expansion is crucial in the AAQG method for maintaining the quality of the generated questions.

4.4.2 Impact on Language Model Answering Capabilities

In table 4, it is evident that the AAQG method results in a significant decrease in pass@1 and pass@5 scores, regardless of whether GPT-3.5-turbo-0125 or CodeLlama-13b-Instruct-hf models are used. This indicates that the AAQG method can effectively disrupt the performance of large language models. Notably, when using GPT-3.5-turbo-0125, the pass@1 score decreased by as much as 42 points.

The table also shows that as the difficulty level increases, the proportion of decline in pass scores decreases. This indicates that packaging simpler problems tends to be more effective, as harder problems are inherently more challenging for LLMs to solve. Conse-

Difficulty	Easy		Medium		Medium_Hard	
	TIS(↑)	KS(↑)	TIS(↑)	KS(↑)	TIS(↑)	KS(↑)
Baseline	0	0.761	0	0.735	0	0.74
AAQG	0.667	0.821	0.662	0.822	0.662	0.824
w/o Skill Type Expansion	0.655	0.752	0.656	0.749	0.656	0.758
w/o Retrieve	0.661	0.751	0.664	0.754	0.661	0.757

Table 3: For evaluating problem quality, the performance of the AAQG method is compared with the baseline and variations of the AAQG method where different components are removed, across different difficulty levels of generated problems.

Method	Model	Easy		Medium		Medium_Hard	
		Pass@1(↓)	Pass@5(↓)	Pass@1(↓)	Pass@5(↓)	Pass@1(↓)	Pass@5(↓)
Baseline	GPT-3.5-turbo-0125	64	-	32	-	36	-
AAQG	GPT-3.5-turbo-0125	22	-	25	-	28	-
Baseline	CodeLlama-13b-Instruct-hf	5.4	13.89	4.8	14.21	1.8	8.11
AAQG	CodeLlama-13b-Instruct-hf	2.0	8.67	1.4	7.0	1.2	5.56

Table 4: The performance of code generation pass rates across different models and difficulty levels.

quently, packaging higher-difficulty problems results in less noticeable improvements.

5 Conclusion

we are the first to explore the task of algorithmic problem generation. We introduced the Automated Algorithmic Question Generation (AAQG) method, which effectively enhances the efficiency of algorithm problem creation. Our experiments demonstrate that problems generated using AAQG can significantly disrupt the code generation capabilities of large language models. Specifically, under the GPT-3.5-turbo-0125 setting, the pass@1 performance decreased by up to 42 points, while in the CodeLlama-13b-Instruct-hf setting, pass@1 dropped by up to 3.4 points and pass@5 decreased by up to 7.21 points.

We also introduced two novel automated evaluation metrics for assessing the quality of generated problems. These metrics demonstrate that problems generated using the AAQG method can maintain the same core essence as the input problems while exhibiting variations.

References

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. arXiv preprint arXiv:2108.07732.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, et al. 2021. Measuring coding challenge competence with apps. arXiv preprint arXiv:2105.09938.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023a. Starcoder: may the source be with you! arXiv preprint arXiv:2305.06161.

Rongao Li, Jie Fu, Bo-Wen Zhang, Tao Huang, Zhihong Sun, Chen Lyu, Guang Liu, Zhi Jin, and Ge Li. 2023b. Taco: Topics in algorithmic code generation dataset. arXiv preprint arXiv:2312.14852.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. Science, 378(6624):1092–1097.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81.

Multi-Linguality Multi-Functionality Multi-Granularity. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. Codegen: An open large language model for code with multi-turn program synthesis. arXiv preprint arXiv:2203.13474.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950.

A Data

The TACO(Li et al., 2023b) dataset, which stands for Topics in Algorithmic COde generation, is primarily designed for the code generation task. This dataset collects problems from various coding platforms such as LeetCode, Codewars, etc. The distribution of the dataset is shown in table 5 . We believe that its rich data is also well-suited for the task of algorithmic problem generation.

Dataset	Train	Test	All
# of Questions	25443	1000	26443

Table 5: This is the data distribution of the TACO dataset.

Table 6 lists the data distribution across all platforms in the TACO dataset, excluding records where the skill_types field is empty, the solution is empty, or the difficulty is labeled as UNKNOWN_DIFFICULTY.

Source	Train	Test	All	Filiter
codeforces	8193	576	8769	4264
aizu	2151	0	2151	0
geeksforgeeks	2680	0	2680	1965
codewars	2460	55	2515	646
kattis	1236	0	1236	0
codechef	3352	278	3630	1268
hackerearth	2390	45	2435	229
atcoder	1440	0	1440	0
leetcode	777	0	777	646
hackerrank	764	46	810	261

Table 6: This Table lists the data distribution for each platform in the TACO dataset.

B Code Generation Detail

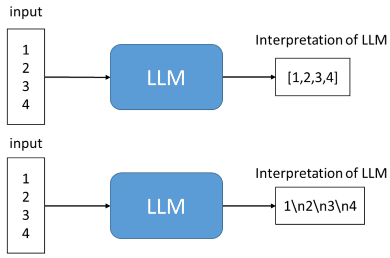


Figure 5: Interpretation of input in LLM: In the same input, the LLM may interpret the data differently, leading to variations in syntax in the output.

During the code generation process, we also use prompting to have the model generate unit test snippets. These snippets are used to guide the model on how to apply the inputs to the generated functions. In this stage of data processing, special cases such as spaces in test data may lead to different interpretations by the model, requiring various data processing methods.

As shown in Figure 5, the LLM may interpret the same input in two different ways, leading to variations in the generated code based on its interpretation. The LLM will process the input according to the format it has interpreted. Therefore, we need to ask the LLM to specify the format of the input that should be provided to its function.

C Pass Rate

As shown in formula eq. (11), where m represents all questions, n indicates the number of attempts the LLM makes to generate code for a single question, and c represents the number of code snippets among these n that pass the unit tests, i.e., the number of correct snippets out of n.

In this formula, when $n = 1$, it directly considers the total number of correct answers. For $n > 1$, it calculates the probability of getting at least one correct answer out of n attempts.

$$pass@k = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}} \tag{11}$$

When using the GPT-3.5-turbo-0125 model, due to cost considerations, it was only run once per attempt, while other models were allowed 10 attempts.

D Ablation Study

D.1 Impact on Language Model Answering Capabilities

In table 7, we used GPT-3.5-turbo-0125 as our code generation model and employed pass@1 as the evaluation metric. We compared different settings of AAQG with the baseline. The results show a significant decrease in scores for AAQG settings compared to the baseline. Among the various settings, the "without Skill Type Expansion" configuration performed best in the Medium_Hard difficulty level, surpassing AAQG by 4 points. However, considering all three difficulty levels, AAQG still demonstrates the best overall performance.

We also observed that in the absence of the retrieve component, the average score across all difficulty levels decreased by 7 points. This indicates that additional relevant articles serve effectively as interference noise in the LLM and can significantly impact the model's ability to generate correct answers.

In tables 8 to 10, we use the same settings as in table 7 but replace the code generation model with CodeLlama-13b-Instruct-hf. This table shows that under different settings of AAQG, there is a noticeable decrease in pass@1 and pass@5 scores compared to the baseline.

In the Easy and Medium difficulty levels, as shown in tables 8 and 9, although the performance in pass@1 and pass@5 is best without Retrieve, the pass@10 scores drop to 55.56% of the original baseline. This suggests that in this setting, the decline in LLM's answering ability may not be due to increased difficulty, but rather due to a lack of sufficient information in most of the generated problems, making many of them unsolvable. This indicates that relying solely on LLM-generated application scenarios can lead to information gaps, resulting in incomplete and unsolvable problems. Therefore, it is crucial to use reference articles to assist in problem generation.

In Medium_Hard difficulty problems, as shown in table 10, although the model performs best without using Retrieve, achieving results equivalent to the baseline, the pass@1 and pass@5 scores show an increase compared to the baseline. This indicates that while reference articles can provide complete information to aid in problem generation, this step may also introduce noise, which can lead to a decline in the LLM's answering ability.

Across all three difficulty levels, it is evident that the complete AAQG setup yields the best overall performance. With only a slight decrease in pass@10 scores, AAQG effectively reduces pass@1 and pass@5 scores. This indicates that problems generated using the AAQG method can significantly interfere with the language model's answering ability.

Method	Easy	Medium	Medium_Hard
Baseline	64	32	36
AAQG	22	25	28
w/o Skill Type Expansion	38	32	24
w/o Retrieve	34	36	26

Table 7: For different difficulty levels and methods, we use the GPT-3.5-turbo-0125 model and employ pass@1 as the evaluation metric to measure the pass rate performance of the generated problems.

Method	Easy		
	Pass@1 (↓)	Pass@5 (↓)	Pass@10 (↑)
Baseline	5.40	13.89	18.0
AAQG	2.0 (-62.96%)	8.67 (-37.58%)	14.0 (-22.22%)
w/o Skill Type Expansion	2.6 (-51.85%)	9.5 (-31.61%)	14.0 (-22.22%)
w/o Retrieve	0.8 (-85.19%)	4.0 (-71.2%)	8.0 (-55.56%)

Table 8: For the Easy difficulty level, we evaluated the model's pass rate performance using different methods. We employed the CodeLlama-13b-Instruct-hf model and used pass@k as the evaluation metric to measure the impact of generated problems on the model's pass rate. The numbers in parentheses next to the pass@k scores indicate the percentage decrease in performance compared to the baseline.

Method	Medium		
	Pass@1 (↓)	Pass@5 (↓)	Pass@10 (↑)
Baseline	4.80	14.21	20.0
AAQG	1.4 (-70.83%)	7.0 (-50.74%)	14.0 (-30%)
w/o Skill Type Expansion	1.6 (-66.67%)	7.54 (-46.8%)	14.0 (-30%)
w/o Retrieve	1.2 (-75%)	6.0 (-57.78%)	12.0 (-40%)

Table 9: For the Medium difficulty level, we evaluated the model's pass rate performance using different methods. We employed the CodeLlama-13b-Instruct-hf model and used pass@k as the evaluation metric to measure the impact of generated problems on the model's pass rate. The numbers in parentheses next to the pass@k scores indicate the percentage decrease in performance compared to the baseline.

Method	Medium_Hard		
	Pass@1 (↓)	Pass@5 (↓)	Pass@10 (↑)
Baseline	1.80	8.11	14.0
AAQG	1.2 (-33.33%)	5.56 (-31.44%)	10.0 (-28.57%)
w/o Skill Type Expansion	1.6 (-11.11%)	6.39 (-21.21%)	10.0 (-28.57%)
w/o Retrieve	2.0 (+11.11%)	8.67 (+6.91%)	14.0 (0%)

Table 10: For the Medium_Hard difficulty level, we evaluated the model's pass rate performance using different methods. We employed the CodeLlama-13b-Instruct-hf model and used pass@k as the evaluation metric to measure the impact of generated problems on the model's pass rate. The numbers in parentheses next to the pass@k scores indicate the percentage decrease in performance compared to the baseline.

Course Stage Recognition for Online Course Recordings Using Spoken Language Understanding

Yi-Ting Yuan, Ke-Ching Hong
Chingshin Academy,
Taipei, Taiwan
{yiting9510,anniehung1030}@gmail.com

Yu-Hsiang Wang *
Department of Electrical Engineering,
National Taiwan University,
Taipei, Taiwan
r11921057@ntu.edu.tw

Abstract

This study investigates models for course stage recognition, a novel task in Spoken Language Understanding (SLU) aimed at segmenting classroom recordings into distinct instructional phases. Two approaches are evaluated: an end-to-end SLU model based on the WavLM base+ speech encoder, and a multistage SLU method integrating Whisper for Automatic Speech Recognition and ChatGPT 4o for Natural Language Understanding. The study compares the performance of these models to explore stage recognition without relying on intermediate text representations. Results indicate that the multistage approach excels in fine-grained classification across five stages—Opening, Lecture, Break, Conclusion, and Others—but is outperformed by the end-to-end model in distinguishing the Lecture stage. The findings suggest that a speech-language model capable of performing in-context learning directly on speech data could further enhance the accuracy of course stage recognition.

Keywords: Course Stage Recognition, Spoken Language Understanding, Speech Model, Large Language Model

1 Introduction

With the rapid advancement of artificial intelligence(AI), speech recognition and large language models (LLMs) have significantly reduced the cost of human-computer interaction, leading to widespread applications in various fields, including education. AI has been applied to assist children's learning (Okur et al., 2023), automatically assess students' attention (Parambil et al., 2022), and analyze classroom discourse to enhance teaching quality (Wang et al., 2024). Although deep learning models have been successfully used in these domains, AI applications for segmenting classroom activities based on content to support teaching organizations mostly rely on traditional machine learning models (Wang et al., 2014; Donnelly et al., 2016a,b).

Classroom activity segmentation can be achieved with a multistage Spoken Language Understanding (SLU) model, using Whisper Large V3 (Radford et al., 2023) for ASR and ChatGPT, the gpt-4o version [1] for NLU. Whisper's noise resilience and multilingual support make it ideal for classroom recordings, while ChatGPT can often interpret correct meanings despite ASR errors. This combination reduces the impact of ASR errors on NLU and enables training-free classroom activity segmentation. However, the reliance on text limits its effectiveness for low-resource or unwritten languages.

To explore alternatives to the multistage SLU approach for segmenting classroom activities, we propose the task of course stage recognition. The goal is to segment long classroom recordings into five categories: Opening, Lecture, Break, Conclusion, and Others.

We developed an end-to-end SLU model using the self-supervised learning speech encoder WavLM base+ (Chen et al., 2022) to extract speech features, followed by Convolutional Neural Networks to reduce the length of the speech features sequence, and Bidirectional Long Short-Term Memory (Schuster and Paliwal, 1997) to predict the classroom stage category for each time frame. This study compares the performance of end-to-end

*Corresponding authors

[1]https://platform.openai.com/docs/models/gpt-4o

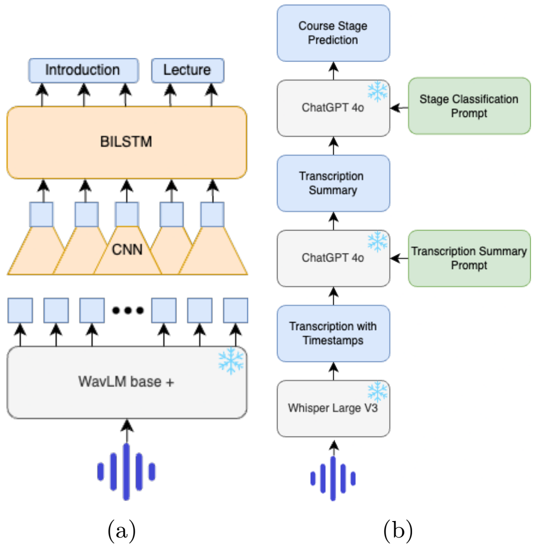


Figure 1: The proposed end-to-end and multistage spoken language understanding (SLU) model. (a) The end-to-end model uses WavLM (Chen et al., 2022) for sequence labeling. (b) The multistage model uses Whisper Large V3 (Radford et al., 2023) for ASR and ChatGPT4-o[2] for NLU.

and multistage SLU models in the classroom stage recognition task. Figure 1 illustrates the model architecture of both SLU models. We summarize the contributions of this study in the following:

- We propose course stage recognition, a novel SLU task designed for long audio recordings.
- We analyze the advantages, limitations, and trade-offs between multistage and end-to-end approaches in the context of course stage recognition.
- To enhance transparency and promote further research, we publicly release the dataset and code used in this study at `https://github.com/yiting9510/Course-Stage-Recognition`.

2 Related Works

In previous studies aimed at improving teaching quality, researchers employed random forest models to classify three types of classroom activities: teacher lecturing, whole class discussion, and student group work, using audio data (Wang et al., 2014). Other studies applied Naïve Bayes classifiers on audio (Donnelly et al., 2016a) and multi-sensor (Donnelly et al., 2016b) data to recognize five instructional segments: question and answer, procedures and directions, supervised seatwork, small group work, and lecture. Our research leverages a more advanced deep learning model to analyze online course audio, categorizing it into five segments: Opening, Lecture, Break, Conclusion, and Others.

Course stage recognition is built upon Spoken Language Understanding (SLU), with two primary approaches: a multistage approach (Bastianelli et al., 2020) that separates Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU), and an end-to-end approach (Wang et al., 2023; Arora et al., 2024). While the multistage method allows independent ASR and NLU training, it is vulnerable to ASR errors in noisy environments like classrooms (Schlotterbeck et al., 2022). The end-to-end approach, although potentially could have better performance, often faces limited training data, especially in specialized domains such as education.

SLU tasks have traditionally used Long Short-Term Memory (Schmidhuber et al., 1997; Schuster and Paliwal, 1997) (LSTM) and encoder-decoder architectures (Sutskever, 2014; Wu, 2016; Chiu et al., 2018), with recent advances in self-attention mechanisms (Vaswani, 2017) enhancing models' ability to capture dependencies within input sequences. These advancements led to the development of large language models (LLMs), such as GPT (Radford et al., 2018, 2019; Brown, 2020), which have strong capabilities in Natural Language Processing (NLP). Speech models built on transformer architectures, like Wav2Vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022), use self-supervised learning (SSL) to learn to extract robust speech features applicable to various speech tasks (wen Yang et al., 2021; Tsai et al., 2022). Whisper (Radford et al., 2023) is an encoder-decoder ASR transformer-based model trained on large annotated datasets, demonstrating strong ASR performance across languages and robust to noisy conditions.

Given the limited labeled data and potential noise in our online course audio, we leveraged pre-trained models and off-the-shelf LLM services to enhance performance. Specifically, we compared two models: a multistage design using Whisper large v3 for ASR and with ChatGPT 4o for NLU (He and Garner, 2023), and an end-to-end model combining WavLM base with CNN and LSTM layers. We evaluated the performance of these models in classroom stage recognition and discussed potential areas for improvement.

3 Method

In this study, we propose a novel task of course stage recognition and have constructed a dataset for this purpose. We developed an end-to-end spoken language understanding (SLU) model based on WavLM and compared it to a multistage SLU model combining Whisper and ChatGPT. We analyzed the performance of these systems in the context of course stage classification. Below, we describe the dataset preparation process and then explain the design of the two SLU models.

3.1 Data Preparation

To create a dataset for course stage recognition, we focused on Mandarin online teaching courses in Taiwan, such as recordings of school lessons for junior high and high school students, as well as videos from tutoring centers. We searched for suitable course recordings on online video platforms and public course websites, extracted the audio, and manually analyzed the content to label different course stages.

Based on the framework outlined by (Davis, 2009), courses can be divided into nine stages: Introduction, Opening, Lecture, Presentation, Break, Transition, Conclusion, Summary, and Others. We used this definition as a reference but merged several similar stages, ultimately classifying the course stages into the following five categories: **Opening** includes teacher greetings and pre-class reminders, which are either unrelated to the main content or provide recaps of previous lessons. **Lecture** covers all content related to the course topic, including guided exercises. **Break** refers to silent periods after the teacher announces a break, as well as off-topic conversations. **Conclusion** summarizes what was covered in the lesson, previews the next session, and includes farewells. **Others** encompasses any other periods, such as time before the class officially begins, after it ends, or interruptions due to technical issues.

Additionally, we defined a simplified classroom stage recognition task, in which courses are categorized into only two stages: Lecture and Others. In this setting, the original Opening, Lecture, and Conclusion stages are grouped under the broad category of Lecture, while Break and any other non-instructional periods are grouped into the Others category.

We manually collected and annotated the recordings according to these standards, labeling each time segment according to its corresponding classroom stage. We ensured that each time segment in the course had only one assigned stage. In total, we collected 65 course recordings and manually annotated the classroom stages. The statistical data of the dataset is shown in Table 1. Notably, the average course length was 15 minutes, which required the method to handle long audio files effectively. The proportion of Lecture stage recordings was 87.8%, significantly higher than the other four stages, necessitating methods capable of addressing extreme class imbalance.

Statistic Description	**Value**
Total Recordings	65
Total Duration (hrs)	90.67
Average Length (min)	15.15
Stage Duration Percentages	
Opening Stage %	2.50%
Lecture Stage %	87.80%
Break Stage %	2.80%
Conclusion Stage %	2.60%
Other Stage %	4.30%

Table 1: Course Recording Statistics

Our observations also revealed that not all five stages were present in every course, and transitions between stages did not always follow fixed patterns. This indicates that methods must be able to analyze contextual information to accurately perform classroom stage recognition.

For the testing dataset, a pre-selected set of 10 recordings was used to ensure coverage of all stages. The remaining 55 recordings are split into training and validation sets at an approximate 8:2 ratio based on their total duration. Two different training-validation splits were randomly sampled, and the average score across these splits was used as the final performance metric.

3.2 Model Design

3.2.1 End-to-End SLU Model

The proposed architecture integrates a Transformer-based Self-Supervised Learning (SSL) speech encoder, Convolutional Neural Networks (CNN), and Bidirectional Long Short-Term Memory (BiLSTM) layers, addressing the task as a sequence labeling problem.

Figure 1a illustrates the end-to-end SLU model structure. The SSL speech encoder, specifically the WavLM Base+ model pre-trained on large-scale speech data, is chosen due to its proven ability to extract robust speech features that generalize well across various speech processing tasks. These pre-trained models have shown superior performances (wen Yang et al., 2021; Tsai et al., 2022; Feng et al., 2023), even when the encoder remains frozen, making them well-suited for the course stage recognition task where labeled data may be scarce. Since the repesentation extracted by the speech encoder is too long for efficient training in stage classification, a CNN is applied to further reduce the sequence length, followed by a BiLSTM layer that classifies each time frame into its corresponding course stage.

In this architecture, the SSL encoder remains frozen, and only the CNN and BiLSTM layers are trained. Experiments were conducted with 2-stage configurations: a detailed 5-stage classification and a simplified 2-stage classification. Due to the computational cost of Transformer-based models increasing rapidly with input length, the maximum input duration is limited to 30 seconds. For longer audio files, a sliding window approach is used to segment them into 30-second chunks with a 10-second overlap.

To address the issue of data imbalance, particularly the overrepresentation of the ”Lecture” stage, several data augmentation techniques were applied. In addition to downsampling the ”Lecture” segments to match the second most frequent class and upsampling non-Lecture segments until their total number reached approximately one-third of the total ”Lecture” segments, we also employed augmentation techniques such as TimeStretch and Gaussian noise. TimeStretch was used to slightly alter the speed of the audio without affecting its pitch, while Gaussian noise was added to enhance robustness against noise in the input data. These augmentations helped improve model generalization and performance, especially in cases where training data was limited.

For model training, cross-entropy loss was used as the loss function. To evaluate model performance, the F1 score was chosen as the primary metric due to its ability to balance precision and recall, especially in imbalanced datasets. The F1 score was computed for each class and then aggregated using either macro-averaging, where all classes are treated equally, or micro-averaging, which in this case is equivalent to accuracy. This provided a comprehensive measure of the model’s classification performance.

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{1}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{2}$$

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{3}$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} F1_c \tag{4}$$

During testing, the model’s predictions are converted into time-based segments, with start time, end time, and the corresponding stage. Both predictions and ground truth are then converted into one-dimensional arrays with 1,000 frames per second, allowing precise alignment. The F1 score is calculated based on these frame-level vectors to evaluate the model’s performance.

3.2.2 Multistage SLU Model

We developed a multistage SLU model combining Whisper large v3 and ChatGPT 4o for classroom stage classification, as illustrated in

Model	Macro-F1 score	Micro-F1 score
End-to-End	0.24559/0.88203	0.82984/0.97111
End-to-End w/ augmentation	0.33477/0.85778	0.84194/0.96629
End-to-End w/ augmentation & undersampling	0.30493/0.88434	0.84299/0.97160
End-to-End w/ augmentation & oversampling	0.39199/**0.91195**	0.85136/**0.97736**
Multistage	**0.49196**/0.74734	0.**85295**/0.94295

Table 2: Comparison of End-to-End model(WavLM base+ with CNN and BiLSTM) and Multistage model(Whisper Large v3 and ChatGPT 4o) performance for class stage recognition performance, where performance under 5 stage (left) and 2 stage (right) settings are shown.

Figure 1b. The model operates in two stages: first, Whisper large v3 performs automatic speech recognition (ASR), converting audio into text with timestamps. We use WhisperX (Bain et al., 2023) for more accurate long-form transcription and to reduce hallucination.

In the second stage, ChatGPT 4o processes the transcribed text for natural language understanding (NLU). It analyzes the transcription to infer classroom stages. Given the length of classroom sessions, ChatGPT 4o uses a two-pass process: first, summarizing chunks of transcription (up to 30 minutes each) to reduce the length, then analyzing the summaries to classify each time segment into its respective classroom stage.

The key benefit of this approach is the use of powerful pre-trained models that require no additional training. With appropriate prompts, the system can accurately predict stages on test data, even when ASR errors are present in the transcription. Performance is evaluated by comparing the predicted stage start/end times with the ground truth using the F1 score.

4 Experimental Results

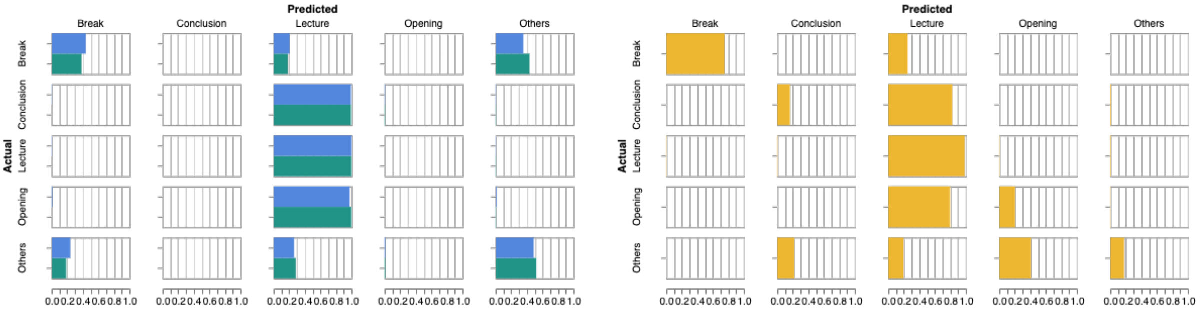
4.1 Model Performance Evaluation

As shown in Table 2, for the 5-stage course classification, the multistage model achieves the highest Macro-F1 score (0.49196) and slightly outperforms the end-to-end models in Micro-F1 score (0.85295 vs. 0.85136). This indicates the multistage model handles class imbalance better, particularly for less frequent categories, resulting in a superior Macro-F1 score. Among the end-to-end models, the one using augmentation and oversampling performs best, with a Macro-F1 score of 0.39199 and Micro-F1 score of 0.85136, though it still lags behind the multistage model in handling imbalanced data.

In the 2-stage classification, both models improve significantly. The end-to-end model with augmentation and oversampling achieves the highest Micro-F1 (0.97736) and Macro-F1 (0.91195) scores, outperforming the multistage model. This result reflects the simpler task's reduced complexity, where the end-to-end model excels by focusing on the two main categories, ”Lecture” and ”Others.” While the multistage model performs decently with a Micro-F1 of 0.94295, it shows a larger gap in Macro-F1 (0.74734), highlighting its less effective handling of the simplified task.

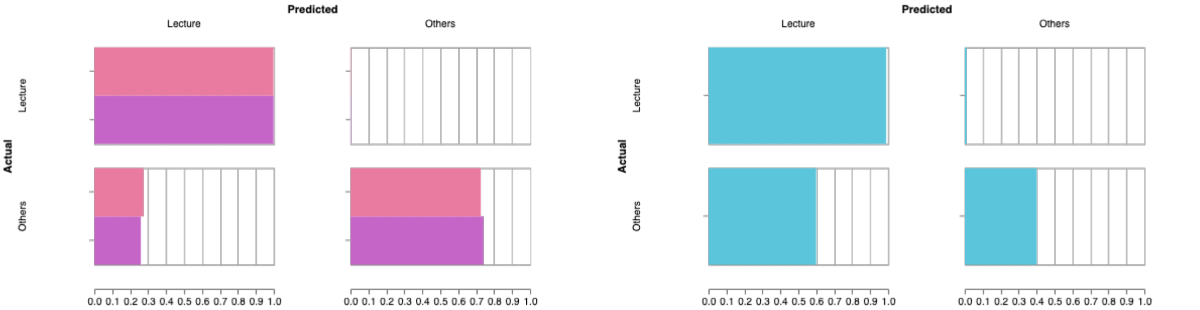
Confusion matrices (Figures 2a and 2b) reveal that in the 5-stage classification, the end-to-end model tends to overpredict the ”Lecture” category, leading to an imbalanced performance, while the multistage model distributes predictions more evenly, contributing to its higher Macro-F1 score. However, in the 2-stage classification, the end-to-end model performs better, reducing prediction imbalance (Figures 3a and 3b).

In summary, each model demonstrates distinct advantages. The end-to-end model performs exceptionally well in the 2-stage task, achieving the highest Micro-F1 and Macro-F1 scores. On the other hand, the multistage model shows superior performance in the more complex 5-stage task, particularly in handling class imbalances. However, both models exhibit limitations: the end-to-end model faces challenges in distinguishing between stages with similar characteristics, while the multistage model risks losing important information, which could impair its ability to accurately recognize certain categories.



(a) (b)

Figure 2: 5-stage recognition Confusion matrix of (a) End-to-end model with data augmentation and oversampling (b) multistage model



(a) (b)

Figure 3: 2-stage recognition Confusion matrix of (a) End-to-end model with data augmentation and oversampling (b) multistage model

4.2 Applications and Impacts

The rise of online educational resources has led to a lack of proper segmentation in video and audio materials, mainly due to the high cost of manual annotation. Course stage recognition addresses this by helping students quickly find relevant content, improving their learning experience. For educators, it offers insights into course structure, enabling better content organization. Educational platforms benefit from more accessible and streamlined content. Our research also advances textless spoken language understanding (SLU), particularly for low-resource and unwritten languages, promoting broader access to educational resources for underrepresented language communities.

4.3 Limitation and Future Work

The end-to-end SLU model in this study faces the challenge of having too short a context window, making it difficult to capture long-term dependencies, which results in an inability to differentiate between similar course stages. On the other hand, multistage models, constrained by their modular design, are prone to losing information, making certain stages harder to recognize. Additionally, relying on ChatGPT for NLU in multistage SLU raises privacy concerns for certain applications.

Future work includes developing an end-to-end SLU model capable of in-context learning. This could be achieved by incorporating a Speech Language Model (SLM), as suggested in recent work (Hsu et al., 2023). The goal would be to use trainable prompts, enabling the SLM (Lakhotia et al., 2021; Kharitonov et al., 2021) to perform SLU on the entire classroom recording, while preserving rich speech information. However, the main challenge is the high computational cost of handling long input sequences.

5 Conclusion

This paper introduces course stage recognition, a novel SLU task aimed at segmenting course content using audio. We propose two models: an end-to-end model based on WavLM and a multistage SLU model using Whisper for transcription and ChatGPT for text understanding. Experimental results demonstrate that both approaches show promising capabilities but have limitations. The end-to-end model can recognize some of the rarer stages but struggles with distinguishing other similar stages, while the multistage model effectively differentiates stages through text analysis but performs worse than the end-to-end model in identifying some of the rarer stages. These results highlight the challenges of course stage recognition. Future work includes developing a speech-language model with in-context learning on speech data to improve performance. We have made our dataset and code publicly available to encourage further research.

Declaration of the Use of Generative AI and AI-assisted Technologies in Writing

During the preparation of this paper, the author(s) used ChatGPT for writing improvement. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Siddhant Arora, Hayato Futami, Jee-weon Jung, Yifan Peng, Roshan Sharma, Yosuke Kashiwagi, Emiru Tsunoo, Karen Livescu, and Shinji Watanabe. 2024. Universlu: Universal spoken language understanding for diverse tasks with natural language instructions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2754–2774.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package. *arXiv preprint arXiv:2011.13205*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4774–4778. IEEE.

Barbara Gross Davis. 2009. *Tools for teaching*. Jossey-Bass.

Patrick J Donnelly, Nathan Blanchard, Borhan Samei, Andrew M Olney, Xiaoyi Sun, Brooke Ward, Sean Kelly, Martin Nystran, and Sidney K D'Mello. 2016a. Automatic teacher modeling from live classroom audio. In *Proceedings of the 2016 conference on user modeling adaptation and personalization*, pages 45–53.

Patrick J Donnelly, Nathaniel Blanchard, Borhan Samei, Andrew M Olney, Xiaoyi Sun, Brooke Ward, Sean Kelly, Martin Nystrand, and Sidney K D'Mello. 2016b. Multi-sensor modeling of teacher instructional segments in live classrooms. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 177–184.

Tzu-hsun Feng, Annie Dong, Ching-Feng Yeh, Shu-wen Yang, Tzu-Quan Lin, Jiatong Shi, Kai-Wei Chang, Zili Huang, Haibin Wu, Xuankai Chang, et al. 2023. Superb@ slt 2022: Challenge on generalization and efficiency of self-supervised speech representation learning. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1096–1103. IEEE.

Mutian He and Philip N Garner. 2023. Can chatgpt detect intent? evaluating large language models for spoken language understanding. *arXiv preprint arXiv:2305.13512*.

Ming-Hao Hsu, Kai-Wei Chang, Shang-Wen Li, and Hung-yi Lee. 2023. An exploration of in-context learning for speech language model. *arXiv preprint arXiv:2310.12477*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, et al. 2021. Text-free prosody-aware generative spoken language modeling. *arXiv preprint arXiv:2109.03264*.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Eda Okur, Roddy Fuentes Alba, Saurav Sahay, and Lama Nachman. 2023. Inspecting spoken language understanding from kids for basic math learning at home. *arXiv preprint arXiv:2306.00482*.

Medha Mohan Ambali Parambil, Luqman Ali, Fady Alnajjar, and Munkhjargal Gochoo. 2022. Smart classroom: A deep learning approach towards attention assessment through class behavior detection. In *2022 Advances in Science and Engineering Technology International Conferences (ASET)*, pages 1–6. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

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

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

Danner Schlotterbeck, Abelino Jiménez, Roberto Araya, Daniela Caballero, Pablo Uribe, and Johan Van der Molen Moris. 2022. "teacher, can you say it again?" improving automatic speech recognition performance over classroom environments with limited data. In *International Conference on Artificial Intelligence in Education*, pages 269–280. Springer.

Jürgen Schmidhuber, Sepp Hochreiter, et al. 1997. Long short-term memory. *Neural Comput*, 9(8):1735–1780.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

I Sutskever. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-Wen Yang, Shuyan Dong, Andy T. Liu, Cheng-I Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2022. SUPERB-SG: enhanced speech processing universal performance benchmark for semantic and generative capabilities. In *ACL (1)*, pages 8479–8492. Association for Computational Linguistics.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Deliang Wang, Yang Tao, and Gaowei Chen. 2024. Artificial intelligence in classroom discourse: A systematic review of the past decade. *International Journal of Educational Research*, 123:102275.

Minghan Wang, Yinglu Li, Jiaxin Guo, Xiaosong Qiao, Zongyao Li, Hengchao Shang, Daimeng Wei, Shimin Tao, Min Zhang, and Hao Yang. 2023. Whislu: End-to-end spoken language understanding with whisper. In *Proc. Interspeech*, volume 2023, pages 770–774.

Zuowei Wang, Xingyu Pan, Kevin F Miller, and Kai S Cortina. 2014. Automatic classification of activities in classroom discourse. *Computers & Education*, 78:115–123.

Yonghui Wu. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Speaker Change Detection Using Ensemble Prediction in Conversations

Qian-Bei Hong
Department of Electrical Engineering
Southern Taiwan University of Science and Technology
qbhong75@gmail.com

Chung-Hsien Wu
Department of Computer Science and Information Engineering
National Cheng Kung University
chunghsienwu@gmail.com

Yeou-Jiunn Chen
Department of Electrical Engineering
Southern Taiwan University of Science and Technology
chenyj@stust.edu.tw

Abstract

Speaker change detection (SCD) is an important technique for separating sentences from different speakers in a conversation. This study proposes a novel approach that utilizes ensemble prediction for SCD. First, a contour predictor is created to predict the frame-based speaker change probability contour for a speech segment. Next, the highest SCD probability for each frame among the sequential segments is selected as the contour value, where each specific frame appears in sequential segments with different timestamps. Finally, the speaker change boundaries are determined using a threshold. In the experiments, the proposed SCD model was evaluated on the 2000 NIST Speaker Recognition Evaluation corpus and the AMI meeting corpus. The NIST corpus was used to train the baseline and proposed models for SCD evaluation. The proposed ensemble prediction achieved the best performance and improved the precision of speaker change boundaries. In addition, the AMI corpus was used to evaluate the effects of out-of-domain prediction. The experiment shows that even though domain mismatch greatly affected the SCD performance, the proposed ensemble prediction, which considers the prediction probabilities of a change boundary from sequential segments achieved improved detection results.

Keywords: Speaker change detection, ensemble prediction, sequential segments

1 Introduction

Speaker change detection (SCD) is an important technique for separating sentences from different speakers in a conversation to understand the relationships between contexts. SCD has been widely used in the automatic speech recognition (ASR) tasks as in (S. Kumar et al., 2024; J. Wu et al., 2023; L. Sarı et al., 2019). It is adopted to split the signal into several speaker-specific speech segments before ASR to avoid transcription error of sequential words presented from different speakers. On the other hand, SCD is an important step for the speaker diarization task (X. Anguera et al., 2012; K. VijayKumar and R.R. Rao, 2023; W. Xia et al., 2022), in which the conversation signal needs to use uniform segmentation to extract the speaker embeddings and cluster them according to speaker features. If a speech segment contains the utterance from more than one speaker, then using SCD to discard the multi-speaker segments before extracting speaker embedding can effectively improve the performance of speaker clustering.

Traditional SCD methods used distance metric such as Bayesian information criterion (BIC) (S. Chen and P.S. Gopalakrishnan, 1998), Kullback-Leibler (KL) divergence (J. E. Rougui et al., 2006), etc. for speaker change detection. However, these traditional techniques often perform poorly due to

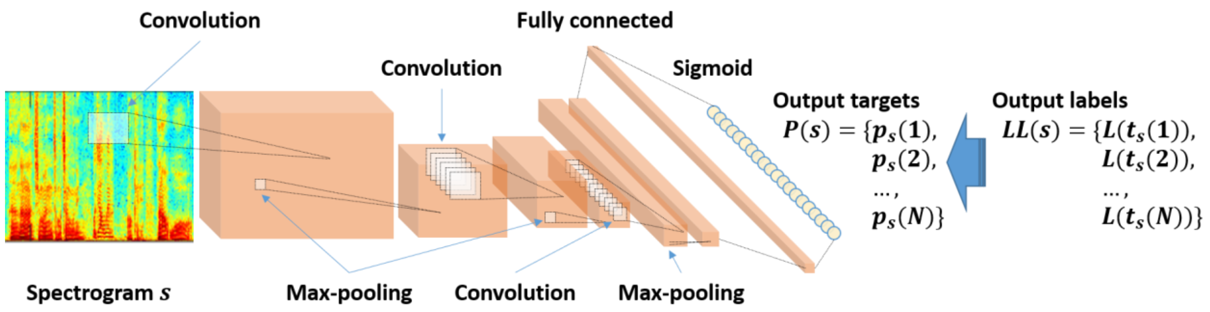


Figure 1. The contour prediction architecture for SCD task.

the challenges associated with processing high-dimensional features in complex multi-speaker conversations. In the past few years, deep neural network (DNN)-based model has been successfully applied to the SCD task (L. Mateju et al., 2019; V. Gupta, 2015; R. Wang et al., 2017), and the performance was better than the traditional methods. (M. Hrúz and Z. Zajíc, 2017) proposed a convolutional neural network (CNN) model with fuzzy label to predict the speaker change boundaries, which has been applied in the diarization systems (Z. Zhou et al., 2018; Z. Zajíc et al., 2018). Fuzzy labelling considers the distance between the detected boundary and the real boundary. The fuzzy labels are used for model training and can effectively improve the detection accuracy. However, in most studies, each speaker change probability is only determined once by the model. Accordingly, the predicted probability is easily affected by different energy distributions and degrades the accuracy of speaker change detection. Thus, this study tries to utilize the ensemble prediction approach to enhance the SCD performance based on the combination of different weights and prediction results.

Ensemble is to integrate different prediction results for performance improvement. The common approach is integrated from multiple models (J. Yi et al., 2017), in which the major point is to predict the output from multiple models with different weights. However, using multiple models for prediction will increase the time cost. Therefore, this study proposes a novel ensemble approach based on multi-label prediction, which requires only one model for prediction.

In summary, this paper uses multi-label prediction to predict the probabilities of each speaker change boundary. The main contributions are as follows: (1) Each potential speaker change boundary can be predicted with multiple values from different interconnection weights, and these probabilities of the same speaker change boundary are integrated to obtain the final prediction results. (2) Using multi-label prediction for segment-based spectrogram can increase the flexibility of frame-level prediction.

2 Speaker Change Contour Prediction

In (M. Hrúz and Z. Zajíc, 2017), a CNN-based SCD model was proposed for speaker diarization, in which the spectrogram of the speech signal is used to predict the speaker change probability of each frame. According to fuzzy labelling definition, the predicted probability is based on the distance between the time point of the current frame and the time point of the real speaker change boundary as follows.

$$L(t) = max\left(0, -\frac{\tau \cdot min_i(|t - b_i|)}{\lambda} + 1\right) \quad (1)$$

where t is the time point of the frame, b_i is the time point of the i-th speaker change boundary closest to t, τ is the shift time between two adjacent frames and $\lambda = 0.6$ is the tolerance. As shown in equation (1), the prediction probability is determined by considering the closest speaker change boundary and the probability of each time point. Thus, the predicted results are easily affected by other magnitudes around the speaker change boundary and decreases the performance of SCD.

2.1 Multi-label Prediction for SCD

In this study, a multi-label prediction approach to SCD is proposed. Figure 1 shows the procedure for multi-label SCD. The multi-label prediction is defined as frame-level prediction for an input

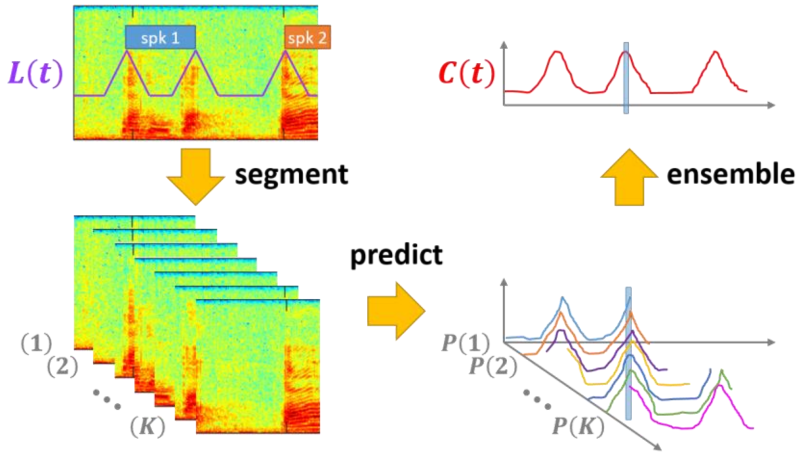


Figure 2. The procedure for the ensemble of prediction probabilities from sequential segments.

Algorithm 1. Boundary merging for SCD.

Input: Contour vector $C \in \mathbb{R}^T$ and change threshold θ

Output: Change time points $Q \in \mathbb{R}^J$

Initial $Q = \emptyset$
Let $\delta = 0.5$ be the onset of a boundary in second.
Let $\tau = 0.01$ be the shift time (in seconds) between two adjacent frames.
$\rho = round(\delta/\tau)$
for $t = 1$ to T **do**
  if $C(t) > \theta$ **then**
    $u = 0, v = 0$
    for $r = t - \rho$ to $t + \rho$ **do**
      if $C(r) > u$ **then**
        $u = C(r), v = r$
      end
    end
    if $v > 0$ and v not in Q **then**
      $Q = \{Q, v\}$
    end
  end
end
return Q

speech segment. The label function is defined as follows.

$$LL(s) = \{L(t) : t_s(1) \leq t \leq t_s(N)\} \quad (2)$$

where N is the number of frames in the speech segment, $t_s(1)$ is the starting time point (first frame) of the s-th segment and $t_s(N)$ is the end time point (last frame) of the s-th segment. We can see that a segment-based spectrogram trained through multi-labels obtains frame-level prediction results, and it only needs to be determined once. In this case, multi-label prediction not only keeps the determined level of fuzzy label, but also increases the prediction precision.

2.2 Ensemble of Predicted Results

After all results of the speech frames in the whole speech signal are predicted, if each time point is included in K segments, the speaker change probability for each time point is determined K times. Therefore, the speaker change contour could be estimated by the ensemble of the K predicted results. Figure 2 shows the ensemble process using multiple prediction to estimate a speaker change contour in the conversation. As the speaker change probability at time point t is predicted K times, the max function is used to estimate the contour value.

$$C(t) = \max_{s \in \Theta_t}\big(p_s(t - t_s(1) + 1)\big) \quad (3)$$

where Θ_t is the set of the speech segments in which the time range includes time point t and $p_s(\cdot)$ is the predicted probability of speaker change for the t-th frame.

2.3 Boundary Merging

The speaker change contour can be used to determine when the speaker changes. The simplest way is to determine the speaker change boundaries by a threshold. Assuming that the threshold $\theta = 0.5$, according to the fuzzy labelling, more than one time points of the same speaker change boundary will be alarmed, which means that many incorrect boundaries are found. Thus, we need to merge the boundaries when these boundaries are too close. In

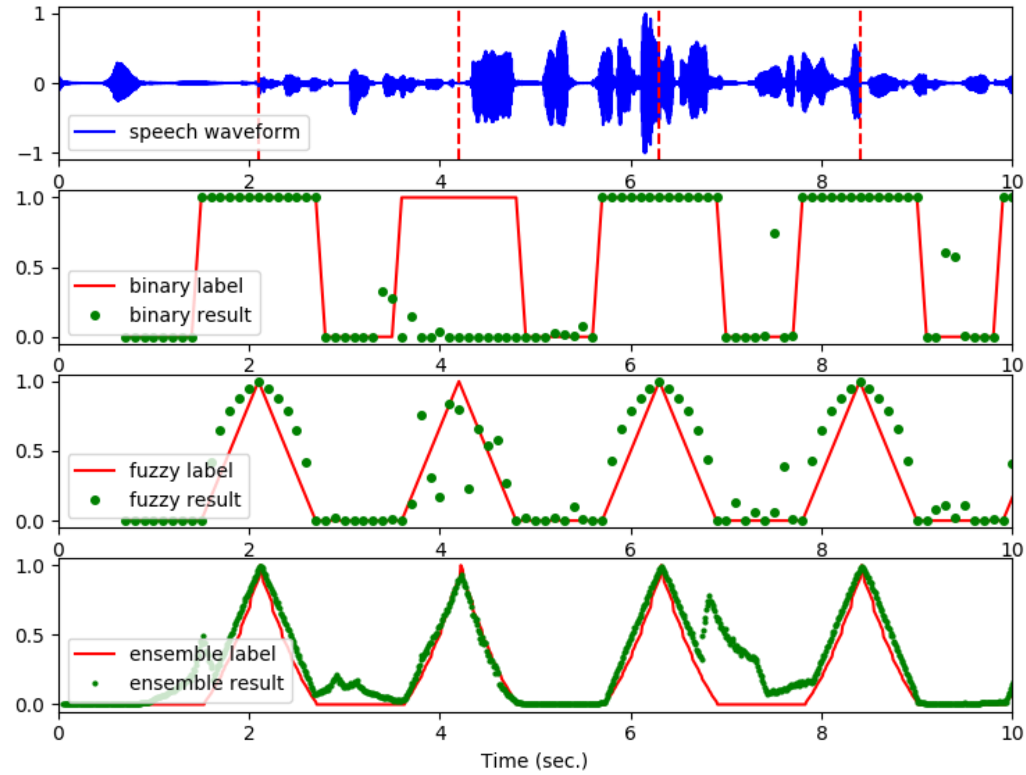


Figure 3. Comparison of the speaker change probability contours obtained from different approaches.

this study, the boundary merging method is shown in Algorithm 1.

3 Experimental Results

3.1 Corpora

This study used two corpora to evaluate the SCD performance and analyzed the effects of domain mismatch.

- **2000 NIST Speaker Recognition Evaluation (LDC2001S97):** This corpus is widely used in speaker diarization experiments. Disk 1 provides audios recorded from 546 females and Disk 2 provides audios recorded from 457 males. We split all audios in these two disks into a sequence of segments of 2.1 second with a time shift of 0.1 second, and we randomly selected and concatenated the segments to generate synthetic conversations for training and testing. This study randomly selected a total of five hours of segments to generate a synthetic conversation for SCD model training and used the randomly selected segments of one hour to generate a synthetic conversation for in-domain data analysis.
- **AMI Meeting Corpus:** The corpus contains 100 hours of meeting recordings. This study utilized the boundary labels of 16 meeting recordings (A. O. T. Hogg et al., 2019) to evaluate the SCD performance in real applications and the effects of out-of-domain data.

3.2 Experimental Setup

The baseline SCD models and the proposed SCD model were trained based on the same backbone architecture as in (M. Hrúz and Z. Zajíc, 2017). The input features were 13-dimentional Mel-frequency cepstral coefficients (MFCCs) with first and second order derivatives. The spectrogram of a segment (consisting of 137 frames) was extracted from a 25ms window with a stride of 10ms, and a

	Acc. (%)	False Alarm (%)	Hit (%)
Binary	86.29	74.41	40.57
Fuzzy	91.14	**17.13**	86.22
Ensemble (proposed)	**92.97**	17.37	**86.29**

Table 1. Results of SCD on the NIST corpus.

Recording Name	Binary		Fuzzy		Ensemble (proposed)	
	Hit (%)	Std. (sec.)	Hit (%)	Std. (sec.)	Hit (%)	Std. (sec.)
ES2004a	72.80	0.2892	62.40	0.2500	75.20	0.2706
ES2004b	63.64	0.2696	57.14	0.2731	76.19	0.2908
ES2004c	76.26	0.2868	65.66	0.2856	80.30	0.3008
ES2004d	60.94	0.2637	68.67	0.2828	72.96	0.2994
IS1009a	72.09	0.2524	81.40	0.3212	76.74	0.2914
IS1009b	76.37	0.2731	75.27	0.2852	73.63	0.2926
IS1009c	69.75	0.2688	69.14	0.2968	74.07	0.2959
IS1009d	81.36	0.2794	66.10	0.2778	79.66	0.2794
EN2002a	66.90	0.2834	68.99	0.2727	73.87	0.2931
EN2002b	62.23	0.2913	59.86	0.2766	70.07	0.2870
EN2002c	60.31	0.2801	59.79	0.2832	67.24	0.3054
EN2002d	69.93	0.2859	65.54	0.2692	75.00	0.3073
TS3003a	85.71	0.2202	47.62	0.2799	80.95	0.3112
TS3003b	84.33	0.2566	63.13	0.3140	73.73	0.3055
TS3003c	78.41	0.2478	60.98	0.3109	77.27	0.3070
TS3003d	75.63	0.2493	65.55	0.3038	78.15	0.3025
Average	72.29	**0.2686**	69.83	0.2864	**75.31**	0.2962

Table 2. Results of SCD on the AMI meeting corpus (domain mismatch).

shift of 0.1 seconds was applied to obtain a sequence of segment-based spectrogram of the input speech signal. The two baseline models ($N = 1$) were trained by the binary labels (M. Hrúz and M. Kunešová, 2016) and fuzzy labels (M. Hrúz and Z. Zajíc, 2017). The speaker change probability is predicted based on the spectrogram of each input frame. The proposed model ($N = 137$) outputted a multi-label prediction.

3.3 Evaluation on NIST Corpus

Because the SCD models were trained by the NIST corpus, the SCD performance can be regarded as in-domain performance. In this experiment, the test signal was obtained by concatenating the speech signals, each with 2.1 seconds from a specific speaker, to form a synthetic conversation, and every ground-truth speaker change applied a collar of 0.2 seconds to tolerate the prediction error.

Figure 3 shows the prediction score contours for speaker change, and the red vertical lines in the first subfigure represents the ground-truth boundaries. We can see that the binary model cannot successfully detect the speaker change boundary, and the fuzzy model does not precisely detect the speaker change boundary and will increase the error of SCD. However, using ensemble prediction can not only increase the time precision of change boundaries, but also improve the detection performance. In Table 1, the experimental results show the performance of SCD in the one-hour synthetic conversations, which contain a total of 1,714 boundaries in the conversations. In the boundary detection task (TRUE or FALSE, decided by the threshold θ), the proposed ensemble prediction achieved the best accuracy of 92.97%, while the binary prediction method achieved an accuracy of 86.29%, which had a very high false alarm rate and a low hit rate. This is because the time point of maximum probability for speaker change is not close to the ground-truth boundary. Therefore, if the test data is in-domain, the ensemble prediction not only achieved the best accuracy of boundary classification, but also obtained the precise time points of SCD.

3.4 Evaluation on AMI Corpus

The AMI corpus has been widely used for SCD system evaluation as in (Z. Fan et al., 2022; M. Kunešová and Z. Zajíc, 2023; S. Kumar et al., 2024), and we used this corpus to evaluate the effects of out-of-domain data. (A. O. T. Hogg et al., 2019) utilized pitch tracking to detect the speaker change boundaries in the AMI corpus. In this study, we used the ground-truth boundaries provided in (A. O. T. Hogg et al., 2019) to evaluate our proposed model. Table 2 shows the SCD performance in 16 AMI recordings. As the SCD models were trained by NIST corpus, using AMI corpus as the testing data will cause the domain mismatch problem. Thus, this experiment used a collar of 0.5 seconds as the tolerance of the prediction error and decreased the threshold θ to increase the hit rate. At the same conditions, we can see that even there are domain mismatch problem in the experiments, our proposed ensemble prediction also achieved the best hit rate. It is worth noting that the hit rate of fuzzy prediction is worse than binary prediction. The reason may be that the fuzzy model with high complexity is difficult to keep accurate prediction for the case of domain mismatch. As the ensemble prediction is based on fuzzy labelling, the performance of ensemble-based SCD will be affected more than binary prediction. In the experiments, the ensemble prediction obtained the average standard deviation of 0.296 seconds for detection error which was slightly worse than the baseline models, but the ensemble prediction achieved an average hit rate of 75.31%, significantly outperforming the binary prediction (72.29%) and fuzzy prediction (69.83%). Therefore, even though the domain mismatch greatly affects the SCD performance, the proposed SCD model based on ensemble prediction can further consider the predicted probabilities of a change boundary in the sequential segments and integrate them to obtain the final prediction results.

4 Conclusions

This paper proposes an ensemble prediction approach to SCD. First, a CNN-based SCD model is trained for multi-label prediction. This method can effectively predict the frame-level results from a segment-level spectrogram. Next, considering each boundary can be predicted more than once through the sequential segments, the maximum probability of speaker change of the same boundary was regarded as the final prediction probability. In the experiments, we found that if the data is in-domain, the proposed ensemble prediction achieved the best hit rate and obtained precise time points of SCD. If the data is out-of-domain, the ensemble prediction can further consider the prediction probabilities of a change boundary in the sequential segments to improve the performance of SCD.

References

S. Kumar, S. Madikeri, I. Nigmatulina, E. Villatoro-Tello, P. Motlicek, K. Pandia, S. P. Dubagunta, A. Ganapathiraju. 2024. Multitask speech recognition and speaker change detection for unknown number of speakers. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12592–12596.

J. Wu, Z. Chen, M. Hu, X. Xiao, and J. Li. 2023. Speaker change detection for transformer transducer ASR. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

L. Sarı, S. Thomas, M. Hasegawa-Johnson and M. Picheny. 2019. Pre-training of speaker embeddings for low-latency speaker change detection in broadcast news. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6286–6290.

X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. 2012. Speaker diarization: a review of recent research. *Journal of the IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356-370.

K. VijayKumar and R.R. Rao. 2023. Optimized speaker change detection approach for speaker segmentation towards speaker diarization based on deep learning. *Journal of the Data & Knowledge Engineering*, 144:102121.

W. Xia, H. Lu, Q. Wang, A. Tripathi, Y. Huang, I. L. Moreno, and H. Sak. 2022. Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8077–8081.

S. Chen and P.S. Gopalakrishnan. 1998. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proceedings of the DARPA broadcast news transcription and understanding workshop*, pages 127–132.

J. E. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez. 2006. Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages V–V.

L. Mateju, P. Cerva, and J. Zdansky. 2019. An approach to online speaker change point detection using DNNs and WFSTs. In *Proceedings of the INTERSPEECH*, pages 649–653.

V. Gupta. 2015. Speaker change point detection using deep neural nets. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4420–4424.

R. Wang, M. Gu, L. Li, M. Xu, and T. F. Zheng. 2017. Speaker segmentation using deep speaker vectors for fast speaker change scenarios. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5420–5424.

M. Hrúz and Z. Zajíc. 2017. Convolutional neural network for speaker change detection in telephone speaker diarization system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4945–4949.

Z. Zhou, Y. Zhang, and Z. Duan. 2018. Joint speaker diarization and recognition using convolutional and recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2496–2500.

Z. Zajíc, M. Kunešová, J. Zelinka, and M. Hrúz. 2018. Zcu-ntis speaker diarization system for the dihard 2018 challenge. In *Proceedings of the INTERSPEECH*, pages 2788–2792.

J. Yi, J. Tao, Z. Wen, and Y. Li. 2017. Distilling knowledge from an ensemble of models for punctuation prediction. In *Proceedings of the INTERSPEECH*, pages 2779–2783.

A. O. T. Hogg, C. Evers, and P. A. Naylor. 2019. Speaker change detection using fundamental frequency with application to multi-talker segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5826–5830.

M. Hrúz and M. Kunešová. 2016. Convolutional neural network in the task of speaker change detection. In *Proceedings of the International Conference on Speech and Computer*, pages 191–198.

Z. Fan, L. Dong, M. Cai, Z. Ma, and B. Xu. 2022. Sequence-level speaker change detection with difference-based continuous integrate-and-fire. *Journal of IEEE Signal Processing Letters*, 29:1551-1554.

M. Kunešová and Z. Zajíc. 2023. Multitask detection of speaker changes, overlapping speech and voice activity using wav2vec 2.0. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

頻率跟隨反應：探索聽覺處理與學習的一扇窗
Frequency Following Response: A Window into Auditory Processing and Learning

Enoch Hsin-Ho Huang and Yu Tsao
Research Center for Information Technology
Academia Sinica
Taipei, Taiwan
enoch.huang@citi.sinica.edu.tw yu.tsao@citi.sinica.edu.tw

摘要

頻率跟隨反應 (Frequency-Following Response, FFR) 是聽力學 (Audiology) 及認知神經科學 (Cognitive Neuroscience) 等領域所關注的一種聽覺電生理訊號，它可以幫助理解腦皮質下 (Subcortical) 區域的聽覺處理與學習，而它的特點是和誘發刺激音具有鎖相 (Phase Locking) 關係。FFR 訊號主要源自腦幹裡中腦的下丘 (Inferior Colliculus) 部位的反應，可透過頭皮電極和相關設備測量並記錄。本文以語音訊號處理的角度介紹 FFR 和相關研究，包括聽覺處理與聽覺電生理等背景知識，並介紹人工電子耳 (Cochlear Implant) 與相關的電誘發頻率跟隨反應 (Electrically-evoked Frequency-Following Response, eFFR)，說明其中的挑戰與機會，以提供參考與啓發。

Abstract

The frequency-following response (FFR) is a type of auditory electrophysiological signal that has gained attention in audiology, cognitive neuroscience, and related fields. It can help understand subcortical auditory processing and learning. The FFR signal exhibits a phase-locking relationship with the stimulating sound. It primarily originates from the response of the inferior colliculus (IC) in the midbrain, part of the brainstem, and can be measured and recorded using scalp electrodes and appropriate equipment. This article introduces the FFR from the perspective of speech signal processing and covers background knowledge on auditory processing and auditory electrophysiology. The cochlear implant (CI) and the related electrically-evoked frequency-following response (eFFR) signal are also introduced, highlighting their challenges and opportunities to offer new insights for readers.

關鍵字：聽覺腦幹反應、聽覺電生理、人工電子耳、頻率跟隨反應、語音訊號處理

Keywords: Auditory Brainstem Response, Auditory Electrophysiology, Cochlear Implant (CI), Frequency-Following Response (FFR), Speech Signal Processing

1 緒論

聽覺是人類重要的感官，它使人透過聲音理解所身處的環境，並透過語言與他人溝通。聽覺感官可分爲周邊聽覺系統 (Peripheral Auditory System) 以及中樞聽覺系統 (Central Auditory System)，前者是從外耳到內耳的部分，在醫學上已有相當豐富的認識 (Pickles, 2013)；後者包含從聽神經至大腦聽覺皮質 (Auditory Cortex) 的範圍，其構造解剖相當精細，而聽覺電生理 (Auditory Electrophysiology) 的研究可幫助人一窺聽覺神經處理與學習機制的奧妙 (楊義良等人, 2020)。

聽覺電生理訊號的發展，主要來自於耳鼻喉科與神經科學兩類領域專家的貢獻 (楊義良等人, 2020)。對於耳鼻喉科 (Otorhinolaryngology) 及聽力學 (Audiology) 而言，由刺激音產生的神經訊號稱爲聽覺誘發反應 (Auditory Evoked Response, AER)；而對於神經科 (Neurology) 和神經科學 (Neuroscience) 的角度來說，聽覺誘發電位 (Auditory Evoked Potential, AEP) 的擷取方式是腦電圖 (Electroencephalography, EEG) 的特例 (楊義良等人, 2020)。

在聽覺電生理相關領域中，頻率跟隨反應 (Frequency-Following Response, FFR) 近年來受到的學者們的關注，因爲它可用來觀察聽覺腦幹 (Auditory Brainstem) 的聲音處理機制，特別是聆聽經驗對於聽覺神經系統的影響 (Kraus et al., 2017)。此外 FFR 訊號僅需數個非侵入式 (Noninvasive) 表面電極置於頭部即可量測，相較腦電圖，只需要較少電極、較少頻道的 FFR，因此在實驗步驟上更爲簡便，所需運算與儲存的資料也較少，因此在科學研究與臨床應用上具有發展的優勢。

本文以循序漸進的方式爲具有語音訊號處理等背景的讀者介紹 FFR，內容包括了聽覺系統、聽覺電生理訊號、FFR 訊號的量測及其與聆聽經驗的關係，以及人工電子耳 (Cochlear Impalnt, CI) 的電誘發頻率跟隨反應 (Electrically-evoked Frequency-Following Response, eFFR)。本文除了說明 FFR 對於聽覺腦幹神經可塑性 (Neuro-plasticity) 的影響，所介紹的相關的生物機制也可對機器學習領域的專家學者提供一些參考與啓發。

2 聽覺系統與聽覺電生理

在說明頻率跟隨反應之前，首先介紹與聽覺系統與電生理相關的背景知識。

2.1 聽覺系統與聽覺電生理

聽覺感官可分爲周邊聽覺系統 (耳朵) 和中樞聽覺系統 (神經系統) 兩部分：

2.1.1 周邊聽覺系統

表1說明人耳由外耳、中耳、內耳三部分組成：外耳包括耳殼 (又稱爲耳廓) 和外耳道，主要負責將空氣分子振動爲形式的聲波收集、加強並引導至中耳。中耳的鼓膜接收空氣分子的波動，透過三塊聽小骨 (Ossicles) 運用槓桿原理將機械波傳遞往內耳。內耳包括負責聽覺的耳蝸和負責平衡的前庭系統，螺旋狀的耳蝸中猶如一個分頻濾波器組，將其中以液體傳播的機械波編碼爲聽神經的電訊號。耳蝸的基底膜扮演了重要的角色，它隔開了分別充滿著內淋巴液 (Endolymph) 和外淋巴液 (Perilymph) 的兩個管狀腔室，藉由聲音行進波 (Traveling Wave) 傳遞，帶動基底膜 (Basilar Membrane) 上的毛細胞 (Hair Cells) 的擺動，使得機械波轉換爲神經性的電脈衝訊號的形式，透過聽覺神經路徑 (Auditory Neural Pathway) 傳遞至大腦。因此，聲音的機械波在耳朵中的傳導介質分別是氣體 (外耳)、固體 (中耳)、液體 (內耳)，如表1。

部位	構造	聲音傳導介質
外耳	耳殼、外耳道	氣　體
中耳	鼓膜、聽骨	固　體
內耳	耳蝸、前庭系統	液　體

表 1: 人耳的主要構造

2.1.2 中樞聽覺系統

中樞聽覺系統又稱爲聽覺神經系統 (Auditory Nervous System)，主要部分是位於聽神經和大腦皮質之間的聽覺神經路徑 (Auditory Neural Pathways)。耳蝸將聲音訊號轉換爲神經脈衝的形式，以便在聽覺神經路徑中傳遞。

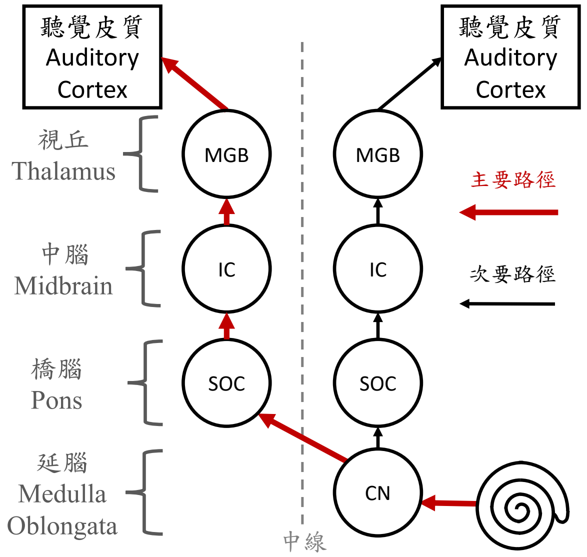


圖 1: 中樞聽覺路徑 (Central Auditory Pathway) (Huang, 2023)，其中的神經核團分別是：CN (Cochlear Nucleus): 耳蝸神經核；SOC (Superior Olivary Complex）: 上橄欖複合體；IC (Inferior Colliculus): 下丘；MGB (Medial Geniculate Body）: 內側膝狀體。自 CN 開始的傳遞路徑，主要是經由對側 SOC 往上，次要路徑則是經由同側。

聽覺神經與前庭神經組成了前庭耳蝸神經 (Vestibulocochlear Nerve)，即第八對腦神經 (Cranial Nerve VIII)，經過數層功能有如中繼站的神經核團 (Nuclei)，最終抵達大腦聽覺皮質。聽神經的傳導包含了兩個方向，分別是往大腦方向傳入 (Afferent) 路徑和反方向的傳出 (Efferent) 路徑，本文僅介紹往大腦方向的聲音傳入路徑。

中樞聽覺傳導路徑如圖1，自耳蝸起，經過同側延腦 (Medulla Oblongata) 中的耳蝸神經核 (Cochlear Nucleus, CN），分爲依照是否穿越中線 (Midline) 分爲對側與同側路徑，其中多數的神經纖維通往對側，因此稱爲主要路徑，而位在較神經纖維的同側則是次要路徑。這兩側的路徑皆經過位在橋腦 (Pons) 的上橄欖複合體 (Superior Olivary Complex）、中腦 (Midbrain) 的下丘 (Inferior Colliculus, IC），此時離開腦幹進到大腦的部位，經過視丘 (Thalamus) 的內側膝狀體 (Medial Geniculate Body, MGB），抵達聽覺皮質 (Pickles, 2013)。

2.2 聽覺電生理訊號

中樞聽覺路徑除了解剖學上對於構造與連接的認識，電生理訊號可幫助理解聽覺神經系統對於誘發訊號的反應。人體中有許多不同種類的

聽覺電生理訊號，學者們發現，依照從刺激音到到聽覺誘發電位出現的時間，可分爲早潛時 (Early Latency)、中潛時 (Middle Latency)、晚潛時 (Late Latency) 反應 (Plourde, 2006)。由於這些客觀的電位反應不會受到個案的意識所影響，因此近年來在臨床與認知相關領域越來越受到重視。

在刺激音後約 15 ms 內誘發的短潛時反應中，聽覺腦幹反應 (Auditory Brainstem Response, ABR) 在臨床上是相廣泛使用的聽力和聽神經的檢查方法，通常採用短聲刺激音進行誘發 (楊義良等人, 2020)。ABR 可客觀地檢查聽力，例如對於無法以言語表達的嬰幼兒，自動 ABR 檢查設備對於聽力異常具有相當不錯的判斷能力，已於納入新生兒聽力篩檢 (陳瑞玲等人, 2015)。此外，ABR 也可以判斷成年人的聆聽閾值，在聽損者的資格或保險失能鑑定等方面廣泛地採用。

有別於以短聲刺激音產生的 ABR，另一種以短語音刺激產生的短潛時反應，可稱爲複合性聽覺腦幹反應 (Complex ABR) 或頻率跟隨反應 (Frequency-Following Response, FFR)，主要是腦皮質下 (Subcortical) 的中腦下丘的反應 (Kraus et al., 2017)。雖然 ABR 有相當成熟的臨床應用，但它的誘發刺激音不是存在於眞實生活中的聲音，因此採用語音刺激更能反應實際的聆聽情形。由於 FFR 在臨床上的應用還不如 ABR 清楚明確，因此有不少學者在深耕與探索中。

3 頻率跟隨反應 (FFR)

FFR 出現已久，近年來重新受到學者們的重視 (Skoe and Kraus, 2010; Kraus et al., 2017; Jacxsens et al., 2024)，介紹如下。

3.1 誘發語音訊號

FFR 實驗以短語音透過耳機播放給受測者的聆聽，例如/da/是常見的刺激音，因爲它包括了子音跟具有週期性音高 (Pitch) 的母音，而且在大多數語言中都有這個音節 (Kraus et al., 2017)。雖然刺激語音的時間長度很短，通常只有數十到數百個毫秒，然而在嚴謹的實驗設計上，一般而言語料的語言和成年人受測個案的母語最好仍要一致。

在英文研究方面，除了以/da/短音作爲刺激語料，有學者研究單純母音/a/、/i/(Aiken and Picton, 2008; Jeng et al., 2024)，或比較/hɛd/、/hid/、/hʊd/比較三個音節中母音的 FFR 表現差異 (Xu et al., 2023)。此外，也有研究是以/ba/、/da/、/ga/等子音爲主 (Hornickel et al., 2009)。

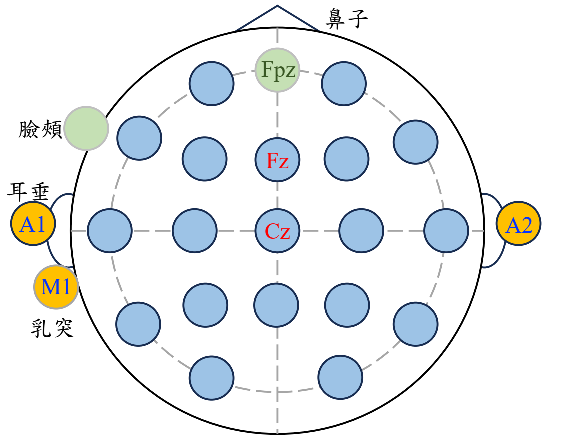


圖 2: 國際 10-20 系統 (19 個淺藍色腦波電極與 2 個參考電極) 與 FFR 和 ABR 的電極位置 (頭頂 Cz 或額頭 Fz；負極：耳垂 A1 或乳突 M1，地極：眉心 Fpz 或臉頰)。淺色圓圈表示不屬於國際 10-20 系統的 ABR 或 FFR 量測電極位置。

在中文的研究方面，台灣曾有學者採用「姨」(一ˊ，/yí/)(Jeng et al., 2011) 語音進行 FFR 實驗，而在美國也曾有針對以中文爲母語者進行實驗，這些個案同時是 11 歲以後才學習英文的雙語者，所採用的刺激語音是「一」的四個聲調，包括「衣」(一，/yī)/)、「姨」(一ˊ，/yí/)、「椅」(一ˇ，/yǐ/)、「易」(一ˋ，/yì/)) (Krishnan et al., 2005)。

3.2 電極位置

FFR 和 ABR 訊號量測的表面電極黏貼位置通常相同，皆是與腦電圖中的電極位置有關。圖2顯示一般腦電圖常用的國際 10-20 系統 (International 10-20 System)，包括 19 個頻道與 2 個參考電極，其中的特定位置與 FFR 和 ABR 的量測位置有關。一般而言，只要三個電極即可表示一個頻道，通常的位置是：正極：Cz 或額頭 Fz；負極：耳垂 A1 或耳後乳突 M1，地極：眉心 Fpz 或臉頰。

3.3 訊號處理

由於 FFR 訊號很微小 ($\leq$ 100 nV)，因此容易受到人體本身和環境中的電訊號干擾 (Jeng et al., 2023)，一種解決的方法是將多次重覆刺激得到的誘發反應進行訊號平均 (Signal Averaging)，例如 1000 次至 6000 次，以消弱背景雜訊的影響 (Skoe and Kraus, 2010)。錄製 FFR 訊號所的儀器，需支援訊號的放大、濾波與錄製等功能 (Jeng et al., 2024)。

爲了改善 FFR 的訊號處理，學者們提出了一些採用機器學習的方法，例如隱藏式馬可夫模型 (Hidden Markov Model, HMM)(Llanos et al., 2017, 2019)、支援向量機 (Support Vec-

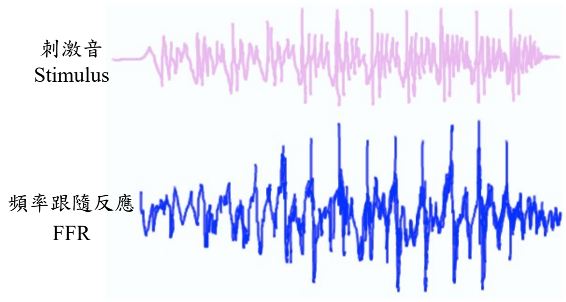


圖 3: 刺激音與理想的 FFR 訊號呈現鎖相 (Phase Locking) 關係 (Kraus et al., 2017)。

tor Machine, SVM) (Xie et al., 2019) 和音源分離非負矩陣分解 (Source Separation Non-negative Matrix Factorization, SSNMF)(Jeng et al., 2023, 2024)，希望更有效且準確地偵測出 FFR 訊號，理想的結果如圖3，可以觀察到刺激音與 FFR 訊號呈現鎖相 (Phase Locking) 關係。

3.4 FFR 與聆聽經驗

研究發現以刺激音誘發的 FFR 反應會受到聆聽經驗的影響，換句話說，以聽覺腦幹的中腦下丘爲主等神經網路是具有學習能力和可塑性的，因此可觀察到 FFR 反應與聲音的學習有關連性，可參考以下語言、音樂和閱讀等例子。

以語言爲例，可以塑造對於聲調的 FFR 反應 (Kraus, 2021)。曾有學者針對不同母語的正常聽力個案進行實驗，結果發現以中文爲母語的受測者相較於英文母語者，在 FFR 波形的音高強度和追蹤的正確性更佳 (Tracking Accuracy)(Jeng et al., 2011)，顯示若從小聆聽聲調語言，聽覺腦幹部位的神經網路已具備音高和語音聲調的學習能力，在進入大腦前已進行過相關聽覺訊號的前處理。

在音樂的聆聽部分，音樂經驗可以增進腦幹的聲音處理。例如，對於有接受音樂訓練的兒童，其 FFR 反應中的訊號雜訊比 (Signal-to-Noise Ratio, SNR) 較未接受音樂訓練的兒童更高 (Skoe and Kraus, 2012)。此外，對於沒有中文語言背景的受測者，其中的音樂家相較於非音樂家的 FFR 訊號，較忠實地反應出對於音高的變化 (Wong et al., 2007)。因此，這些研究都說明了音樂對於聽覺腦幹的可塑性 (Kraus and Chandrasekaran, 2010)。

此外，一個有意思的研究是關於閱讀障礙 (Dyslexia)。對於聽力正常卻患有閱讀障礙的兒童，在聆聽多次相同刺激音實驗中，他們的 FFR 反應會有不一致性，換句話說，即使是播放相同的聲音，但他們每次聽到的會有些不同。研究指出，使用課堂調頻系統 (FM System) 搭配類似助聽器的聆聽輔助裝置 (Listening Aid) 一年後，即使這些聽力正常的兒童不再配戴聽力的情況下，仍可以有效減少他們 FFR 反應上的變異性，進而提高閱讀能力和語音意識 (Speech Awareness)，也證實了這些在個案身上的改變是包括了皮質下的神經性學習結果。(Hornickel et al., 2012; Kraus et al., 2017)。

4 人工電子耳頻率跟隨反應

本節介紹與人工電子耳相關的電誘發頻率跟隨反應 (Electrically-evoked Frequency-Following Response, eFFR)。

4.1 人工電子耳簡介

人工電子耳又稱爲人工耳蝸 (Cochlear Implant, CI)，是目前對重度聽損者最有效的聽覺輔具。由於大多數重度聽損者已無法聽到助聽器所放大的聲音，因此透過手術植入將電極陣列耳蝸內後，可產生特定的電刺激形式 (Stimulating Patterns) 使神經脈衝訊號通過聽覺神經系統來感受到聽力。電子耳是近半個世紀以來在醫學工程領域中相當偉大的發明 (Zeng et al., 2008; Clark, 2015; Wilson, 2019) ，目前全球植入人工電子耳的使用者已超過一百萬人 (Zeng, 2022)，而在台灣的電子耳醫療經驗也相當地成熟 (吳哲民與鄒詠婷, 2015)。雖然電子耳可成功地幫助重度聽損恢復溝通能力，不同電子耳使用者的語音理解度差異 (Variability in Speech Intelligiblity) 相當大 (Wilson, 2019)，能夠呈現個體差異表現的 FFR 訊號或許有機會提供較客觀的分析與解釋。

電子耳的系統架構如圖4，分爲體外和體內兩部分。是一種使用電流刺激的聽力輔具。電子耳在人體外的部分是可以更換與升級的，包括了聲音處理器 (Sound Processor) 和發射線圈。聲波由聲音處理器的麥克風所接收，經過類比數位轉換 (Analog-to-Digital Converter) 和數位訊號處理，以線圈發射無線電訊號至體內，經接收解碼後，由電極產生電流刺激聽神經。爲了改善聆聽效果，電子耳的訊號處理方法在持續地改良中，例如噪音消除 (Noise Reduction)(Lai et al., 2017) 和聲音編碼策略 (Sound Coding Strategy)(Huang et al., 2021, 2024)，隨著深度學習與人工智慧科技的發展不斷前進 (黃心和、吳炤民與曹昱, 2023)。在電子耳的體內部分，在接收到體外的無線電後，以植入在耳蝸內的電陣列 (例如最普遍的設計是 22 個電極)，依照音高位置的不同刺

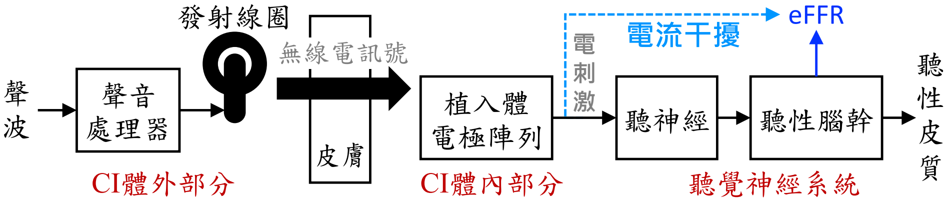


圖 4: 人工電子耳系統 (體外與體內部分)、聽覺神經系統、電誘發頻率跟隨反應 (eFFR) 訊號，以及 CI 電刺激對於 eFFR 的電生理訊號造成干擾的示意圖。

激聽神經，所產生的動作電位 (Action Potentials) 經過聽覺腦幹的神經路徑達到大腦的聽覺皮質而感受到聲音。

4.2 電誘發頻率跟隨反應

CI 使用者的聽覺腦幹也有 FFR 反應，不同的是刺激的訊號來源是電流，因此稱為 eFFR (Electrical FFR) 反應，有別於一般以聲音做為刺激訊號的 aFFR(Acoustic FFR)(Venâncio et al., 2022)。

電誘發頻率跟隨反應 (eFFR) 可用於評估人工電子耳使用者的聽覺腦幹處理，幫助理解神經系統對於聽力個體差異的影響，由於 CI 是以電刺激 (Electric Stimulation) 而非聲刺激 (Acoustic Stimulation) 產生反應，故訊號處理的電干擾挑戰更大，如圖4中藍色虛線表示電流干擾，可能改變原本的 eFFR 波形。因此有學者提出模板刪減法 (Template Subtraction) 方法來消除電刺激干擾 (Gransier et al., 2024)，或是嘗試國際 10-20 標準的其他電極位置 (Van Canneyt et al., 2017)。此外，也有學者提出讓正常聽力個案聆聽 CI 聲碼器 (Vocoder) 以及低通濾波器 (Lowpass Filter) 所結合的模擬語音，以評估電聲雙模聆聽 (Bimodal) 的效果 (Xu et al., 2023)。

因為相較於一般以聲音誘發的 FFR，eFFR 實驗能更直接的進行電刺激訊號的控制，但電刺激訊號也容易干擾 eFFR 訊號，因此在聽覺電生理訊號處理與聲音編碼與學習的領域既是挑戰也是機會

5 討論與結論

FFR 的研究除了幫助研究人員了解聽覺相關的神經機制，也可了解如何透過聲音訓練幫助聽損者與閱讀障礙者進改善其生活品質，對於聲音感知的生理及心理機制也能促進更深入的認識。由於人工電子耳的電聽覺 (Electric Hearing) 可提供聲聽覺 (Acoustic Hearing) 較無法觀察到的一些面向，值得更多的探索。

FFR 的重要推手是美國西北大學的 Nina Kraus 教授，她和團隊在神經科學與聽力學有許多的貢獻。Kraus 教授曾主編了一本關於 FFR 的工具書《The frequency-following response: A window into human communication》(Kraus et al., 2017)，也撰寫了一本針對一般大眾的科普性質的書籍《Of Sound Mind》（中譯：《大腦這樣「聽」》，天下文化，2022 年）(Kraus, 2021)，其中包括了相當豐富的參考文獻，希望藉此推廣關於 FFR 的知識，使更多人了解甚至投入相關的研究。

在生成式人工智慧 (Generative Artificial Intelligence) 崛起的今日 (Feuerriegel et al., 2024)，FFR 相關研究的發現提醒著我們重新檢視對於現有科技和對聽覺生理 (Auditory Physiology) 的認識。人工神經網路 (Artificial Neural Networks) 日新月異，它們的設計最初是模仿生物神經網路 (Biological Neural Networks)，而此二種神經網路均有值得再思之處 (Hasson et al., 2020)。隨著人工智慧與機器學習等技術快速發展並應用到聽力保健 (Hearing Healthcare) 相關領域 (Lesica et al., 2021) 時，推陳出新的網路架構與仿生設計的初衷似乎漸行漸遠。此外，生理上的聽覺傳導路徑通常被視為剛性的神經訊號傳輸通道，(Pickles, 2013)，然而，研究證據指出聽覺路徑本身即是具有可塑性的神經網路，包括 FFR 相關的研究，顯示相關的神經網路是可以透過聆聽的訓練或經驗而調整輸出至大腦的聽覺神經訊號。從訊號處理的角度而言，透過訓練而增強的前端聽覺路徑，是有助於後續大腦聽覺處理的 (Kraus et al., 2017)，若能透過觀察 FFR 等電生理訊號的客觀結果，為聽損者等人士發展有效改善聽覺神經處理的聽力訓練或復健方法，值得更深入的探索。

本文是從語音訊號處理的角度對於 FFR 的回顧性文章，除了提供聽覺系統和電生理學的背景知識，也介紹相關的研究成果。FFR 不僅能顯示腦幹樞紐位置的下丘的反應，也可呈現個體的差異和學習的影響。除了以聲音刺激的 FFR 反應，人工電子耳使用者對於電刺激的 eFFR 反應也是值得探討的議題。期盼能對於語言、聽力、和聲音訊號處理等專業背景的研究人員帶來一些啓發。

參考文獻

Steven J Aiken and Terence W Picton. 2008. Envelope and spectral frequency-following responses to vowel sounds. *Hearing research*, 245(1-2):35–47.

Graeme M Clark. 2015. The multi-channel cochlear implant: Multi-disciplinary development of electrical stimulation of the cochlea and the resulting clinical benefit. *Hearing research*, 322:4–13.

Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. 2024. Generative ai. *Business & Information Systems Engineering*, 66(1):111–126.

Robin Gransier, Robert P Carlyon, Matthew L Richardson, John C Middlebrooks, and Jan Wouters. 2024. Artifact removal by template subtraction enables recordings of the frequency following response in cochlear-implant users. *Scientific reports*, 14(1):6158.

Uri Hasson, Samuel A Nastase, and Ariel Goldstein. 2020. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434.

Jane Hornickel, Erika Skoe, Trent Nicol, Steven Zecker, and Nina Kraus. 2009. Subcortical differentiation of stop consonants relates to reading and speech-in-noise perception. *Proceedings of the National Academy of Sciences*, 106(31):13022–13027.

Jane Hornickel, Steven G Zecker, Ann R Bradlow, and Nina Kraus. 2012. Assistive listening devices drive neuroplasticity in children with dyslexia. *Proceedings of the National Academy of Sciences*, 109(41):16731–16736.

Enoch Hsin-Ho Huang. 2023. *Investigations of cochlear implant sound coding strategies based on auditory physiology and deep learning*. National Central University, Taoyuan, Taiwan.

Enoch Hsin-Ho Huang, Rong Chao, Yu Tsao, and Chao-Min Wu. 2024. Electrodenet—a deep-learning-based sound coding strategy for cochlear implants. *IEEE Transactions on Cognitive and Developmental Systems*, 16(1):346–357.

Enoch Hsin-Ho Huang, Chao-Min Wu, and Hung-Ching Lin. 2021. Combination and comparison of sound coding strategies using cochlear implant simulation with mandarin speech. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:2407–2416.

Laura Jacxsens, Lana Biot, Carles Escera, Annick Gilles, Emilie Cardon, Vincent Van Rompaey, Willem De Hertogh, and Marc JW Lammers. 2024. Frequency-following responses in sensorineural hearing loss: A systematic review. *Journal of the Association for Research in Otolaryngology*, 25(2):131–147.

Fuh-Cherng Jeng, Hsiung-Kwang Chung, Chia-Der Lin, Brenda Dickman, and Jiong Hu. 2011. Exponential modeling of human frequency-following responses to voice pitch. *International journal of audiology*, 50(9):582–593.

Fuh-Cherng Jeng, Tzu-Hao Lin, Breanna N Hart, Karen Montgomery-Reagan, and Kalyn McDonald. 2023. Non-negative matrix factorization improves the efficiency of recording frequency-following responses in normal-hearing adults and neonates. *International Journal of Audiology*, 62(7):688–698.

Fuh-Cherng Jeng, Katie Matzdorf, Kassy L Hickman, Sydney W Bauer, Amanda E Carriero, Kalyn McDonald, Tzu-Hao Lin, and Ching-Yuan Wang. 2024. Advancing auditory processing by detecting frequency-following responses through a specialized machine learning model. *Perceptual and Motor Skills*, 131(2):417–431.

Nina Kraus. 2021. *Of sound mind: How our brain constructs a meaningful sonic world*. MIT Press.

Nina Kraus, Samira Anderson, and Travis White-Schwoch. 2017. *The frequency-following response: A window into human communication*. Springer.

Nina Kraus and Bharath Chandrasekaran. 2010. Music training for the development of auditory skills. *Nature Reviews Neuroscience*, 11(8):599–605.

Ananthanarayan Krishnan, Yisheng Xu, Jackson Gandour, and Peter Cariani. 2005. Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research*, 25(1):161–168.

Ying-Hui Lai, Fei Chen, Syu-Siang Wang, Xugang Lu, Yu Tsao, and Chin-Hui Lee. 2017. A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation. *IEEE Transactions on Biomedical Engineering*, 64(7):1568–1578.

Nicholas A Lesica, Nishchay Mehta, Joseph G Manjaly, Li Deng, Blake S Wilson, and Fan-Gang Zeng. 2021. Harnessing the power of artificial intelligence to transform hearing healthcare and research. *Nature Machine Intelligence*, 3(10):840–849.

Fernando Llanos, Zilong Xie, and Bharath Chandrasekaran. 2017. Hidden Markov modeling of frequency-following responses to Mandarin lexical tones. *Journal of neuroscience methods*, 291:101–112.

Fernando Llanos, Zilong Xie, and Bharath Chandrasekaran. 2019. Biometric identification of listener identity from frequency following responses to speech. *Journal of neural engineering*, 16(5):056004.

James O. Pickles. 2013. *An introduction to the physiology of hearing*, fourth edition. Brill.

G Plourde. 2006. Auditory evoked potentials. *Best Practice & Research Clinical Anaesthesiology*, 20(1):129–139.

Erika Skoe and Nina Kraus. 2010. Auditory brain stem response to complex sounds: A tutorial. *Ear and hearing*, 31(3):302–324.

Erika Skoe and Nina Kraus. 2012. A little goes a long way: How the adult brain is shaped by musical training in childhood. *Journal of Neuroscience*, 32(34):11507–11510.

Jana Van Canneyt, Michael Hofmann, Tom Francart, and Jan Wouters. 2017. Frequency following responses in cochlear implant users. In *Conference on Implantable Auditory Protheses, Date: 2017/07/16-2017/07/21, Location: Lake Tahoe, USA*.

Leonardo Gleygson Angelo Venâncio, Mariana de Carvalho Leal, Laís Cristine Delgado da Hora, Silvana Maria Sobral Griz, and Lilian Ferreira Muniz. 2022. Frequency-following response (ffr) in cochlear implant users: a systematic review of acquisition parameters, analysis, and outcomes. In *CoDAS*, volume 34, page e20210116. SciELO Brasil.

Blake S Wilson. 2019. The remarkable cochlear implant and possibilities for the next large step forward. *Acoustics Today*, 15(1):53–61.

Patrick CM Wong, Erika Skoe, Nicole M Russo, Tasha Dees, and Nina Kraus. 2007. Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature neuroscience*, 10(4):420–422.

Zilong Xie, Rachel Reetzke, and Bharath Chandrasekaran. 2019. Machine learning approaches to analyze speech-evoked neurophysiological responses. *Journal of Speech, Language, and Hearing Research*, 62(3):587–601.

Can Xu, Fan-Yin Cheng, Sarah Medina, Erica Eng, René Gifford, and Spencer Smith. 2023. Objective discrimination of bimodal speech using frequency following responses. *Hearing research*, 437:108853.

Fan-Gang Zeng. 2022. Celebrating the one millionth cochlear implant. *JASA Express Letters*, 2(7).

Fan-Gang Zeng, Stephen Rebscher, William Harrison, Xiaoan Sun, and Haihong Feng. 2008. Cochlear implants: System design, integration, and evaluation. *IEEE Reviews in Biomedical Engineering*, 1:115–142.

吳哲民與鄒詠婷. 2015. 人工電子耳綜論 (Cochlear implantation)[in Chinese]. 台灣耳鼻喉頭頸外科雜誌, 50(4):197–210.

楊義良等人. 2020. 電生理聽力學 *(Auditory Electrophysiology)[in Chinese]*. 華騰文化，台北市.

陳瑞玲等人. 2015. 台灣實施公費新生兒聽力篩檢之過去, 現況與未來 (The past, present and future of newborn hearing screening program in Taiwan)[in Chinese]. 台灣耳鼻喉頭頸外科雜誌, 50(2):67–73.

黃心和、吳炤民與曹昱. 2023. 人工電子耳聲音訊號處理：通往人工智慧的創新旅程 (Sound processing for cochlear implants: The journey of innovation toward artificial intelligence)[in Chinese]. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 227–232.

Advanced Personal Voice Activity Detection through Attention Score module with Conformer Block and FiLM Layers
通過 Conformer 模塊與特徵線性調製層結合自注意力分數損失進行個人語音活動偵測

Ruei-Xian Chang[1], En-Lun Yu[2], Berlin Chen[2], Shih-Chieh Huang[3] and Jeih-Weih Hung[1]
[1]National Chi Nan University, [2]National Taiwan Normal University, [3]Realtek Semiconductor Corp.
s110352016@ncnu.edu.tw, enlunyu@ntnu.edu.tw, berlin@ntnu.edu.tw,
eric.sc.huang@realtek.com, jwhung@ncnu.edu.tw

摘要

近年來，深度學習已成為全球熱門話題，作為人工智慧的一個重要分支，越來越多學者致力於推動這一領域的發展，語音處理技術在科技的進步下也有了大幅度的進展。語音活動檢測(voice activity detection, VAD)是一種重要的語音預處理步驟，通常位於其他語音處理任務之模型的上游，藉由分隔出待處理的語音段落，可以有效降低所串接之語音處理模型的運算複雜度。在本論文中，我們針對語者特定之個人式 VAD (personal VAD, PVAD) 的模型加以開發改進，在提出的新 PVAD 架構中，我們參照了 AS-pVAD 模型當中的自注意力分數模塊(AS block)、將其整合特徵線性調製層(FiLM layers)以更有效率地整合語者資訊、同時將一般 PVAD 之 LSTM 模型架構以 Conformer block 加以取代。經過這樣的模型改良，實驗結果顯示相較於基礎模型架構，目標語者平均精確度增加將近 5%，在其他各項評估分數也有顯著的提升。

Abstract

Deep learning's growing prominence in artificial intelligence has significantly impacted speech processing. Voice activity detection (VAD) acts as a crucial pre-processing step, isolating speech segments for further processing by downstream models, thereby reducing their computational burden. This paper addresses advancements in speaker-specific personal VAD (PVAD) models. Our proposed PVAD architecture leverages the self-attention score module (AS block) from the AS-pVAD model. To more effectively integrate speaker information, we incorporate feature-wise linear modulation layers (FiLM layers) into the architecture. Additionally, we replace the standard Long Short-Term Memory (LSTM) architecture with a Conformer block for enhanced performance.

Experimental results demonstrate a significant improvement in target speaker accuracy (nearly 5% increase) and other evaluation metrics compared to the baseline model. These findings highlight the effectiveness of the proposed modifications in speaker-specific VAD tasks.

關鍵字：語音活動檢測、自注意力分數模塊、特徵線性調製層
Keywords: voice activity detection, attentive score block, feature-wise transform layers, conformer

1. 緒論 (Introduction)

語音活動偵測(Voice Activity Detection, VAD)(Florian Eyben et al., 2013)是一種常見的語音預處理系統，其檢測語音活動是否為人聲及非人聲，因其過濾非語音的特性，常被作為各種語音相關任務、如自動語音辨識(Automatic Speech Recognition, ASR)(Takenori Yoshimura et al., 2020)、語者驗證(Speaker Verification, SV)(Li Wan et al., 2018)和語者分離(speaker separation)(Quan Wang et al., 2018)等架構的前端，以降低系統的資源消耗。

在實際應用場景中，由於通常存在語者干擾與人聲雜訊 (babble noise) 等具有類似語音成分卻不須被關注的語音，導致 VAD 系統在進行偵測時有較高的誤判率，進而使後端的處理系統對於誤判為語音的訊號進行多餘的處理，產生消耗計算資源（如 CPU、內存和電池）、降低處理速度的問題，這對於資源有限的語音處理系統影響尤其明顯。因此，一個優異的 VAD 系統，應該在實際應用場景中精確地區隔語音與非語音片段、同時，在目標語者與語音接收或處理裝置進行對話時，才觸發下游高功率元件進行下一步處理。為

此，文獻(Shaojin Ding et al., 2020)對音框為單位(frame-wise)的 VAD 的架構衍伸出個人語音活動偵測 (Personal Voice Activity Detection, PVAD)的系統，這裡稱為 PVAD 1.0。相較於 VAD，PVAD 僅識別目標語者的語音、忽略非目標語者與純雜訊的片段、更加關注在於目標語者的語音活動，進一步減少後端計算資源的消耗。PVAD性能的提升，有助於個人化設備和服務（如智能家具和藍芽音控設備及智慧型 3C 產品）的發展與應用。

近年來，有諸多的 PVAD 相關研究成果已被發表，包含提出原始 PVAD 1.0 的研究團隊，也提出了進階版，稱作 PVAD 2.0 (Shaojin Ding et al.,2022)，另外，一個基於注意力得分(attention score)機制的輕量級 PVAD (Attentive Score Personal Voice Activity Detection, AS-pVAD)(Fenting Liu et al.,2024)其效能相當優異，AS-pVAD 內部包含了高效提取語者嵌碼(speaker embedding)的模型，除了以低參數量模型解決現有諸多 PVAD 模型對於語者嵌碼系統的外部需求外，AS-pVAD 藉由兩階段偵測的方式，使整體架構對於未註冊語者(unenrolled speaker) 的語句也能發揮一般的 VAD 之效能。此外，AS-pVAD 模型中也採用了一種創新的計算注意力得分損失函數(Attentive Score)之模組，使整體模型更關注於目標語者相關的聲學特徵，使整體 PVAD 架構對於目標語者之嵌碼內容能學習地更加完整。

儘管 AS-pVAD 的效能優異，我們認為它仍然有幾項待改進的部分：

1. 注意力模組效能會隨語者嵌碼(speaker embdding)的表示能力而有顯著影響，因此，使用表示能力較差的語者嵌碼可能會使模型的學習表現不佳，注意力模組對整體 PVAD 的貢獻幅度有限。

2. 注意力模組僅透過串接來結合聲學特徵及語者嵌碼，我們認為在這兩部分應有更具效率的方式來進行整合。

在本文中，我們提出新的一個 PVAD 架構，為上述原始 AS-pVAD 的潛在問題提出改善的方案，實驗結果表示，我們新提出的 PVAD 架構，能夠在相似的模型參數量時，取得更大效益的結果。

以下是本文之內容安排：第二節簡要回顧了 PVAD 1.0 (Shaojin Ding et al., 2020)及 PVAD 2.0 (Shaojin Ding et al., 2022) 模型及提出的新方法流程，第三節包含了實驗設置與實作細節，第四節包含了實驗結果與討論，最後，第五節則為本文結論。

2. 提出方法 (Proposed Method)

2.1 回顧個人語音偵測與個人語音偵測 2.0

PVAD 1.0 法(Shaojin Ding et al.,2020)先從預訓練的語者驗證模型提取目標語者嵌碼(Speaker Embedding)，再將語音從特徵擷取器(Feature Extractor)提取的對數梅爾器組能量音訊特徵(log Mel-filterbank energies)進行串接，輸入個人語音偵測主架構，最後送進輸入線性層，得到以音框為單位(frame-wise)的目標語者決策機率：

$$\hat{F}_t = [F_t; e^{target}], \quad (1)$$

$$p = PVAD(\hat{F}_t), \quad (2)$$

其中F_t為音框 t 之音訊特徵，e^{target}為目標語者嵌碼，p為 PVAD 輸出之該音框對於到目標語者之機率。

然而，在上述的 PVAD 1.0 架構中，語音特徵透過串接特徵來訓練模型可能不是最佳效益的方法，其進階版 PVAD 2.0 (Shaojin Ding et al.,2022)提出使用特徵線性調製層(Feature-wise Linear Modulation layer, FILM layer)代替串接進行特徵融合。因此，PVAD 2.0 求取目標語者的機率過程如下：

$$\overline{F_t} = LSTM(F_t), \quad (3)$$

$$\widetilde{F_t} = \gamma(e^{target}) \cdot \overline{F_t} + \beta(e^{target}), \quad (4)$$

$$p = L(\widetilde{F_t}), \quad (5)$$

其中，$LSTM(\cdot)$為長短期記憶層操作層、其對於音訊特徵加以編碼而成$\overline{F_t}$、$\gamma(\cdot)$和$\beta(\cdot)$分別是作用於目標語者嵌碼e^{target}之 FiLM layer 的縮放和偏移向量，$L(.)$為一線性層(linear layer)。

2.2 注意力分數模塊(AS)

如前所述，文獻(Fenting Liu et al.,2024)在 PVAD 的框架下提出了注意力分數模塊(Attentive Score Block)，此模塊對於目標語者嵌碼及音訊特徵的串接$\hat{F}_t = [F_t; e^{target}]$（如式

(1)所示），透過兩個卷積層提取兩者的相似度分數：

$$M_t = AS(\hat{F}_t), \quad (6)$$

其中，$AS(.)$為上述兩個卷積層之運算，此相似度分數M_t和$\widehat{F_t}$進行逐元素相乘，獲得加權特徵$F_{AS,t}$：

$$F_{AS,t} = \hat{F}_t \odot M_t。 \quad (7)$$

上述之使用串聯法結合目標語者嵌碼e^{target}及音訊特徵F_t來求取相似度(注意力)分數M_t，可能有幾項潛在缺點：

1. 維度特徵增加；
2. 串聯內容不同的資訊模態使得模型學習能力降低。

2.3 與注意力分數模塊結合特徵線性調製層(FiLM-AT-PVAD)

基於上述之原始注意力分數模塊在輸入上的缺點，本論文提出一個進階版的注意力分數模塊，其流程圖如圖一(a)表示，它首先使用目標語者嵌碼e^{target}，求取一特徵線性調製層(FiLM Layer)，其縮放與偏移向量分別以γ和β表示，此 FiLM 層運作在音訊特徵F_t，得到新特徵

$$\tilde{\mathbf{F}}_t = \mathrm{FiLM}(\mathbf{F}_t) = \gamma(\mathbf{e}^{\mathrm{target}}) \cdot \mathbf{F}_t + \beta(\mathbf{e}^{\mathrm{target}}), \quad (8)$$

$\tilde{\mathbf{F}}_t$ 相當於整合了目標語者嵌碼e^{target}與音訊特徵$\mathbf{F}_t$兩方的資訊，接著，$\tilde{\mathbf{F}}_t$ 用於求取相似度(注意力)分數M_t、進而將M_t與音訊特徵F_t點乘而得到輸出加權特徵$F_{AS,t}$，此$F_{AS,t}$通過一個由數層 LSTM 與全連接層(fully connected layer, FC)構成的分類層(classification layer)，求取各輸入音框對應之目標語者的機率。此 PVAD 整體流程圖如圖一(b)所示，為了敘述簡易起見，我們以將其命名為 **FiLM-AT-PVAD**。其中，我們預設音訊特徵為透過 Feature Extractor 模組產生的對數梅爾器組能量音訊特徵F_t，而分類層中的 LSTM 層數為 2。

針對上述所新提出的 **FiLM-AT-PVAD** 架構，我們同時對其加以變化，探索是否能達到更佳的 PVAD 效能，以下，是我們研發出的四種變形：

1. **FiLM-AT-PVAD$_{L1, L1}$:**
   在 **Feature extractor 模組後多加一層 LSTM，而分類層中的 LSTM 層數由 2 降為 1**，其目的是**藉由 LSTM 層對於音訊特徵加以編碼優化、卻簡化分類層的參數量**，探索在不明顯更動原 FiLM-AT-PVAD 的複雜度前提下，觀察其效能是否能有所提升，我們將其命名為 FiLM-AT-PVAD$_{L1, L1}$, 其下標 L1, L1 分別代表注意力模組前端為 1 層 LSTM 與後端為 1 層 LSTM。
2. **FiLM-AT-PVAD$_{L1, L2}$:**
   在 Feature extractor 模組後多加一層 LSTM，而分類層中的 LSTM 層數維持 2，其目的是**探討額外增添 LSTM 的預處理對於音訊特徵加以優化後**，是否能帶來改善。我們將其命名為 FiLM-AT-PVAD$_{L1, L2}$, 其下標 L1, L2 分別代表注意力模組前端為 1 層 LSTM 與後端為 2 層 LSTM。
3. **FiLM-AT-PVAD$_{C1, L1}$**
   在 **Feature extractor 模組後多加一層 Conformer模塊，而分類層中的LSTM層數由 2 降為 1**，在此 Conformer 模塊(Anmol Gulati et al.,2020)作為音訊特徵的編碼器，Conformer 模塊為自注意力與卷積神經網路的結合，可進行全局自注意力交互學習，處理時間序列任務，其中自注意力機制幫助模型捕捉時間序列資料中的長期依賴關係，而卷積層捕捉局部特徵，前饋模組進一步變換特徵，增加非線性處理能力以增強模型的整體表達能力，有效的捕捉相對偏移的局部相關性。由於 Conformer 層較 LSTM 層精細與複雜，預期它的參與會提升 PVAD 之效能，但會增加運算複雜度。我們將其命名為 FiLM-AT-PVAD$_{C1, L1}$, 其下標 C1, L1 分別代表注意力模組前端為 1 層 Conformer 與後端為 1 層 LSTM。
4. **FiLM-AT-PVAD$_{C1, L2}$**
   在 **Feature extractor 模組後多加一層 Conformer模塊，而分類層中的LSTM層數維持為 2**，因此我們將其命名為 FiLM-AT-PVAD$_{C1, L2}$, 其下標 C1, L2 分別代表注意力模組前端為 1 層 Conformer 與後端為 2 層 LSTM。

這四種 FilM-AT-PVAD 之變形的流程圖分別如圖二(a)(b)與圖三(a)(b)所示。另外，根據上述的命名法，由於原始 FiLM-AT-PVAD 在 Feature extractor 模組後並未增加

任何模塊，而分類層的 LSTM 層數為 2，因此我們將其命名為 FiLM-AT-PVAD$_{\emptyset, L2}$，其下標$\emptyset$, L2 分別代表注意力模組前端無額外模組、後端為 2 層 LSTM。

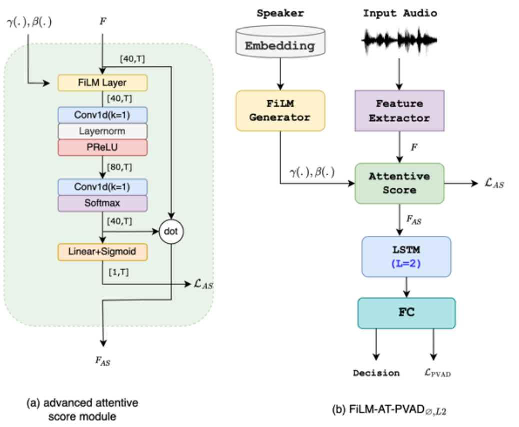


圖 1：(a)進階式注意力分數模塊 (b)原始 FiLM-AT-PVAD (FiLM-AT-PVAD$\emptyset$, L2)架構

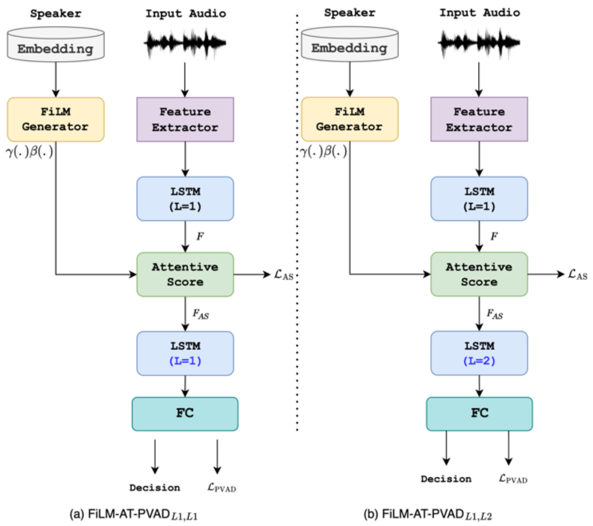


圖 2：(a) FiLM-AT-PVAD$_{L1, L1}$ 架構 (b) FiLM-AT-PVAD$_{L1, L2}$ 架構，二者皆使用一層的 LSTM 作為音訊特徵的編碼器，不同在於後端之 LSTM 層數

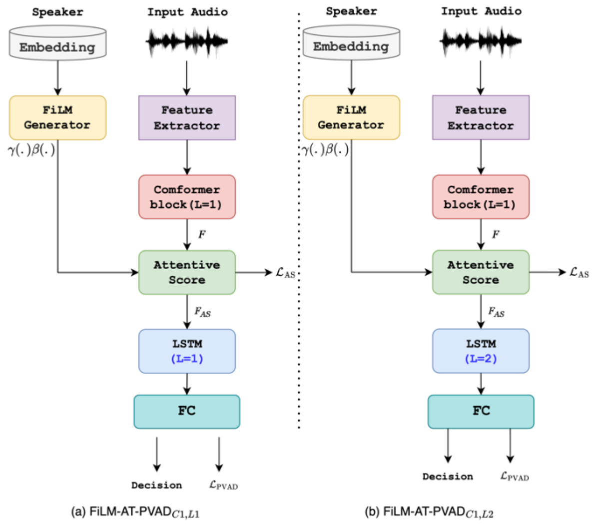


圖 3：(a) FiLM-AT-PVAD$_{C1, L1}$ 架構 (b) FiLM-AT-PVAD$_{C1, L2}$ 架構，二者皆使用一層的 Conformer 作為音訊特徵的編碼器，不同在於後端之 LSTM 層數

2.4 損失函數(Loss function)

為了進行二元分類的PVAD任務，我們使用二元交叉熵(Binary Cross Entropy, BCE)(Zhilu Zhang et al., 2018)作為注意力分數模塊的損失函數之一，其計算方法如下：

$$L_{PVAD} = \frac{1}{T}\sum_{t=0}^{T-1} BCE(y_t, p_t) \qquad (9)$$

上式中，$y_t \in \{0,1\}$和$p_t \in \{0,1\}$分別代表音框 t 的真實標籤值(目標語者為 1，其他為 0)和 PVAD 的預測標籤值，T是總時間框數量。

此外，基於(Fenting Liu et al.,2024)，我們引入了注意力分數模塊損失函數L_{AS}來專門訓練注意力分數模塊。注意力分數模塊損失通過計算真實標籤y_t 與估計的加權分數$\widetilde{m}_t$之間的均方根誤差(Mean Squared Error)損失，促使注意力分數模塊專注於學習跨多模態信息的相似性。

$$\widetilde{m}_t = \text{Sigmoid}(\text{L}(M_t)) \qquad (10)$$

$$L_{AS} = \frac{1}{T}\sum_{t=0}^{T-1}(y_t - \widetilde{m}_t)^2 \qquad (11)$$

公式(10)中的M_t是在公式(6)中顯示的相似度分數，Sigmoid (·)和 L(·)分別表示 Sigmoid 層和線性層操作。最終，將上述兩個損失相加，得到訓練整個 PVAD 模型的總損失：

$$L_{total} = L_{PVAD} + L_{AS}。 \qquad (12)$$

3. 實驗設置 Experiment setup

3.1 實驗數據集

我們採用了 LibriSpeech 訓練集(Vassil Panayotov et al., 2018)來評估各種 PVAD 的效能。該訓練集包含三個子集，總計 960 小時來自 2338 位不同說話者的語音數據：訓練集 train-clean-100 和 train-clean-360 提供了總共 460 小時的乾淨語音，而 train-other-500 提供了 500 小時的雜訊語音。同樣地，LibriSpeech 的測試集包含乾淨和雜訊語音，總計 10 小時來自 73 位語者的語音。為了進行實驗，數據集必須包含多個語者之語音的串接。因此我們首先利用均勻分布、選擇一到三作為語者的數目，並將它們的語音連接起來，之後隨機選擇其中一位語者作為目標語者。在這個過程，我們以 0.2 的比率加入未包含目標語者語音的多語者語音，避免模型過度學習目標語者的特徵。

關於說話者嵌碼，我們為每位說話者隨機挑選語音並輸入預訓練的說話者驗證模型，以生成窗級的 d 向量。這些 d 向量經 L2 正規化並平均，進而產生作為對應整段語句(utterance-based)之目標語者嵌碼的之 d 向量 e^{target}。

此外，為了防止學習過擬合並使模型具有強健性，我們使用了 MTR 的數據增強技術(Chanwoo Kim et al., 2017)，功能是引入了具有不同房間模擬脈衝響應的隨機噪聲源，這種數據增強技術能有效地提高模型對不同噪聲和混響條件的模擬，讓模型能夠在更廣泛的現實環境中表現良好。

3.2 實作細節

我們對語音數據及提取 40 維的梅爾濾波器能量作為原始聲學特徵，其框架大小為 25 毫秒，間隔為 10 毫秒。對於我們所實驗的各種 PVAD 模型的參數設定，詳細介紹如下：

(1) AT-PVAD 基礎模型：
此為使用原始 AS-pVAD 之 AS 模組(以語者嵌碼與音訊特徵為輸入，如式(6)(7))，其輸出再經由兩層 LSTM 與兩層線性層組成的分類層，這兩層 LSTM 各有 64 個單元(unit)。

(2) FiLM-AT-PVAD$_{\emptyset, L2}$：注意力模塊後端的兩層 LSTM 各有 64 個單元(unit)。

(3) FiLM-AT-PVAD$_{L1, L1}$：
注意力模塊前端的一層 LSTM 包含 40 個單元，而後端的一層 LSTM 包含 64 層單元。

(4) FiLM-AT-PVAD$_{L1, L2}$：
注意力模塊前端的一層 LSTM 包含 40 個單元，而後端的兩層 LSTM 各有 64 個單元。

(5) FiLM-AT-PVAD$_{C1, L1}$
注意力模塊前端的一層 Conformer 模塊是由兩個前饋模塊、一個自注意力模塊及一個卷積模塊所組成，輸出 40 個單元，而後端的一層 LSTM 包含 64 個單元。

(6) FiLM-AT-PVAD$_{C1, L2}$
注意力模塊前端的一層 Conformer 模塊是由兩個前饋模塊、一個自注意力模塊及一個卷積模塊所組成，輸出 40 個單元，而後端的兩層 LSTM 各有 64 個單元。

對於整體實驗，我們使用 PyTorch (Adam Paszke et al., 2019)實現了所有模型。在模型訓練中，我們最初使用 Adam 優化器(Diederik P. Kingma et al., 2014)，在第一個訓練周期的學習率為1×10^3，隨後在接下來的訓練周期中將學習率降低到 1×10^5。

4. 實驗結果與討論 Results&Discussions

我們採用了多種不同指標來評估各個 PVAD 模型。其中，平均精度(AP)為目標說話者語音(tss)及非目標語音與非語音(ntss&ns) 精確率-召回率曲線下的面積，平均精度期望值(mAP)則是再對平均精度進行加權平均，數值越高，表示模型在檢測目標語者音框的精確性越高；準確率(Acc.)(%)是正確檢測的音框數與總檢測音框數的比率，反映模型在區分目標語者語音與非目標語者音框方面的準確性。由於我們所採用的數據集所包含的正樣本數少於負樣本，因此我們在使用平均精度期望值時會是一個相當重要的指標，此外，我們還透過考慮模型參數的數量(Par.)來評估模型在資源有限上的適用性。

表 1：各種 PVAD 對應的目標語者平均精度、非目標語者語音平均精度、平均精度期望值、準確率及模型參數量，其中 AT-PVAD baseline、FiLM-AT-PVAD$_{\emptyset, L2}$、FiLM-AT-PVAD$_{L1, L2}$、FiLM-AT-PVAD$_{C1, L2}$ **後端皆有 2 層 LSTM**

model	tts	ntss&ns	mAP	Acc.(%)	Par.(k)
PVAD 1.0	88.4	94.7	92.2	84.34	130.24
PVAD 2.0	90.8	96.0	94.1	86.58	97.60
AT-PVAD 基礎模型	86.8	94.4	91.6	83.79	**84.95**
FiLM-AT-PVAD$_{\emptyset, L2}$	90.3	95.1	93.4	86.08	92.03
FiLM-AT-PVAD$_{L1, L2}$	88.8	94.7	92.5	84.84	105.15
FiLM-AT-PVAD$_{C1, L2}$	**95.9**	**98.3**	**97.5**	**91.99**	131.95

表 2：PVAD 2.0 與兩種較輕量之 PVAD 的評估分數，其中 FiLM-AT-PVAD$_{L1, L1}$、FiLM-AT-PVAD$_{C1, L1}$ **後端皆有 1 層 LSTM**

model	tts	ntss&ns	mAP	Acc.(%)	Par.(k)
PVAD 2.0	90.8	96.0	94.1	86.58	97.60
FiLM-AT-PVAD$_{L1, L1}$	91.2	95.5	94.0	86.74	**71.87**
FiLM-AT-PVAD$_{C1, L1}$	**95.6**	**98.2**	**97.3**	**91.47**	98.67

4.1 比較使用 ASPVAD 提出的方法

表 1 顯示了 PVAD 1.0 (Shaojin Ding et al., 2020)、PVAD2.0 (Shaojin Ding et al., 2022)及 AT-PVAD 基礎模型和提出的三種較複雜之模型（後端有 2 層 LSTM）進行比較，從此表格當中，我們有下列觀察與分析：

1) 三種提出的模型(FiLM-AT-PVAD$_{\emptyset, L2}$、FiLM-AT-PVAD$_{L1, L2}$、FiLM-AT-PVAD$_{C1, L2}$)表現都優於 AT-PVAD 基礎模型，由此可以看出我們所提出之注意力分數模塊（藉由 FiLM 整合語者資訊e^{target}與音訊特徵$\mathbf{F}_t$）優於原始注意力分數模塊（單純串接e^{target}與音訊特徵$\mathbf{F}_t$）。

2) AT-PVAD 基礎模型效果相對較差，而使用同一種（原始）注意力分數模塊之 AS-pVAD 法(Fenting Liu et al.,2024)、根據其文獻之成果，效果極佳，這可能是因為它額外採用了更先進的語者嵌碼提取器（ECAPA-TDNN）和聲學特徵編碼器（TCNN），為注意力分數模塊提供了更優越的輸入特徵。

3) FiLM-AT-PVAD$_{L1, L2}$ 相較於 FiLM-AT-PVAD$_{\emptyset, L2}$ 額外於注意力分數模塊前採用一層 LSTM 作為音訊特徵編碼，效果卻反而變差，然而，當使用一層 Conformer 來作作為音訊特徵編碼的 FiLM-AT-PVAD$_{C1, L2}$，在各項指標上都是最佳的，但由於 Conformer 的複雜性，使其對應的模型參數量最多。

4.2 分析模型簡化後的結果

表 2 顯示了 PVAD2.0 及提出的兩種簡化的模型（FiLM-AT-PVAD$_{L1, L1}$、FiLM-AT-PVAD$_{C1, L1}$）的各項指標，觀察此表、並與表 1 相較，我們觀察到：

1) FiLM-AT-PVAD$_{L1, L1}$ 相對於 FiLM-AT-PVAD$_{L1, L2}$ 少了一層 LSTM，雖然模型變小，但各項指標分數皆提升了許多，其中幾項重要的數據：目標語者平均精度(tts)及準確率(Acc.)超過了 PVAD 2.0(AP:91.2 vs. 90.8, Accuracy:86.74 vs. 比 86.58)，這顯示出多一層 LSTM 於後端的分類器效果不一定比較好，對於注意力分數模塊捕捉到的特徵在經過二層 LSTM 之分類器時可能造成梯度消失的結果。

2) FiLM-AT-PVAD$_{C1, L1}$ 相較於 FiLM-AT-PVAD$_{C1, L2}$ 少了一層 LSTM，效果僅微幅下降，但模型

大小大幅減少（98.67 k vs. 131.949k），這使得它更適合應用在資源有限的設備上，在實際應用中具有顯著優勢。

5. 結論與未來期望 (Conclusion and future works)

本研究拓展新穎之 AS-pVAD 其注意力分數模塊(AS block)的概念，採用特徵線性調製層(FiLM layer)來整合語者嵌碼與音訊特徵，並提出用 Conformer 模塊來對音訊特徵作編碼，而初步實驗證實我們所提出的方法在實際應用中能夠顯著讓 PVAD 模型效果提升，在未來展望上，希望能透過所提出之 FiLM-AT-PVAD$_{C1, L1}$ 模型作為基礎、設計出表現更佳更輕量的 PVAD 模型。

參考文獻 (References)

Florian Eyben, F. Weninger, S. Squartini, and B. Schuller. 2013. *Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies.* 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.

Takenori Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe. 2020. *End-to-End Automatic Speech Recognition Integrated With CTC-Based Voice Activity Detection.* 2020 IEEE International Conference on Acoustics, Speech and Signal Processing.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. *Generalized end-to-end loss for speaker verification.* 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. 2018. *Speaker diarization with LSTM.* 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Shaojin Ding, Quan Wang, Shuo-yiin Chang, Li Wan, Ignacio Lopez Moreno. 2020. *Personal VAD: Speaker-Conditioned Voice Activity Detection.* The Speaker and Language Recognition Workshop

Fenting Liu, Feifei Xiong, Yiya Hao, Kechenying Zhou, Chenhui Zhang, Jinwei Feng. 2024. *AS-pVAD: A Frame-Wise Personalized Voice Activity Detection Network with Attentive Score Loss.* ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Shaojin Ding, Rajeev Rikhye, Qiao Liang, Yanzhang He, Quan Wang, Arun Narayanan, Tom O'Malley, Ian McGraw. 2022. *Personal VAD 2.0: Optimizing Personal Voice Activity Detection for On-Device Speech Recognition.* Interspeech.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang. 2020. *Conformer: Convolution-augmented Transformer for Speech Recognition.* Interspeech.

Jiawei Xiao and Peng Lu. 2024. *A Hybrid Model of Conformer and LSTM for Ocean Wave Height Prediction.* Applied Sciences.

Zhilu Zhang and M. R. Sabuncu. 2018. *Generalized cross entropy loss for training deep neural networks with noisy labels.* Proceedings of the 32nd International Conference on Neural Information Processing Systems, Curran Associates Inc.

Vassil Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. *Librispeech: An asr corpus based on public domain audio books.* 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara Sainath, and Michiel Bacchiani. 2017. *Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home. Proc.* Interspeech 2017, pp. 379–383.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, Soumith Chintala. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library.* Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc.

Diederik P. Kingma, Jimmy Ba. 2014. *Adam: A Method for Stochastic Optimization.* arXiv preprint arXiv:1412.6980

Design and Development of a Speech Assistive Device for Esophageal Speakers
食道語者語音輔助裝置之設計與開發

Yan-Zhi Chen, Yen-Ting Lin, Bo-Sen Liang, Ding-Lian Chen, Chen-Yu Chiang
National Taipei University, New Taipei City, Taiwan
oscar900517@gmail.com, d26923050@gmail.com, benson20030603@gmail.com,
shadow449515@gmail.com, cychiang@mail.ntpu.edu.tw

摘要

食道語者因疾病或手術等原因失去聲帶，無法像大多數人一樣使用聲帶發音，而是使用食道振動來取代聲帶振動。然而，食道語在發聲過程中需要頻繁吞氣，產生大量雜音和爆音，使得一般人較難以理解。爲了解決這一問題，我們開發了一個優化食道語的系統。該系統通過分析食道語的頻譜特性，並且利用機器學習技術來過濾雜音並抑制爆音。我們將食道語分爲四類：爆音（burst, B）、正常語音（voice, V）、雜音（noise, N）和靜音（silent, S）。這個系統的主要目的是在使用較少計算資源的前提下，消除食道語中的雜音並抑制爆音，從實驗結果顯示來看確實有顯著提高食道語的可理解性。

Abstract

Esophageal speakers lose their vocal cords due to conditions such as illness or surgery, making it impossible for them to use their vocal cords to produce sound like most people. Instead, they use esophageal vibrations to replace vocal cord vibrations. However, during the phonation process of esophageal speech, frequent air swallowing is required, which produces a lot of noise and burst sounds, making it difficult for the average person to understand. To address this issue, we developed a system to enhance esophageal speech. This system analyzes the spectral characteristics of esophageal speech and uses machine learning techniques to filter out noise and suppress burst sounds. We categorize esophageal speech into four types: bursts (B), normal voice (V), noise (N), and silence (S). The primary goal of this system is to eliminate noise and suppress burst sounds in esophageal speech while using fewer computational resources. The experimental results show a significant improvement in the intelligibility of esophageal speech.

關鍵字：語音轉換、語音增強、食道語

Keywords: Voice conversion, Speech enhancement, Esophageal speech

1 Introduction

1.1 研究背景及方向

食道語 (Kaye et al., 2017) 是利用食道内的空氣振動來產生聲音。食道語者通過吞咽或吸入空氣進入食道，然後利用食道和咽喉部的肌肉控制這些空氣的排出，在此過程中形成振動並產生聲音。這些聲音再通過口腔和鼻腔共鳴，產生可理解的語音，許多人因爲各種原因而失去了聲帶，不得不使用食道語與他人進行溝通。然而，學習並熟練掌握食道語需要努力且長久的練習，並不是一朝一夕可以精通的。食道語的發聲過程中，由於需要頻繁吞咽空氣來維持聲音的產生，這會導致顯著的噪音問題。吞咽動作本身會產生低頻的吞咽聲，而空氣在進入食道和從食道排出時，也會產生明顯的進氣和排氣聲。這些聲音，尤其是進氣聲，往往是突然且不規則的，並且可能帶有爆發性的音質，這會增加語音中的背景雜音，進而影響語音的清晰度和可理解性。這種噪音不僅讓聽者難以理解語音内容，還會增加交流的困難度，形成的雜音也更容易導致他人難以理解。

本研究錄製了一個食道語者的語音，使用 VoiceBank-2023 語料庫 (Su et al., 2023) 裡面總計 105 句的文本，總錄製時長約 18 分鐘。使用 VoiceBank-2023 語料庫是因爲語料庫有經過特別的設計，照順序累積語句，可以比較以較少的語句來包含中文的聲母和韻母，即是在語句檔案較少的情況下，也能得到更好的效果。

從食道語者的語音中可以觀察到，食道語者在發音時一開始的聲音振幅偏大，這是因爲中文大部分的音節從 consonant（子音）開始，加上他們只能利用食道内的空氣壓力來擠

出聲音，這導致接在子音後面的 vowel（母音）或韻母音量較小，可能在時域上產生了「遮蔽效應」，使母音相較子音顯得較不清晰。我們觀察到無聲子音如ㄐ、ㄑ、ㄒ、ㄓ、ㄔ、ㄕ、ㄗ、ㄘ、ㄙ等容易會產生爆音，因此我們把上述歸類在爆音（B）這個標籤，而無聲子音像是ㄅ、ㄆ、ㄈ、ㄉ、ㄊ、ㄍ、ㄎ、ㄏ雖是無聲子音但是根據觀察，發生爆音的概率較低，因此我們把上述類別和有聲子音ㄇ、ㄋ、ㄌ、ㄖ及母音歸類在正常語音（V）的標籤裡面，只降低容易爆音的無聲子音振幅並預期會有顯著改善食道語的清晰度。Table 1 是我們統計 105 句語音檔案，判斷爆音的標準爲振幅超過我們所設定的閥值 0.7 時，即被判定爲爆音。語音振幅超過 0.7 時對應的標籤結果如下:

Label	count	Probability (%)
B	1603	87.452%
V	220	12.002%
S	1	0.055%
N	9	0.491%

Table 1: 爆音對應標籤的發生機率

根據上表可以看出，我們這樣的分類方法是可行的，也能發現歸類在爆音 (B) 類別中的無聲子音確實容易導致爆音的現象。這說明無聲子音對於食道語者發音時，更容易觸發爆音的產生。透過這樣的分類，我們可以更清楚地了解各類聲音在語音處理中的行爲，並進一步優化語音處理系統的性能。

「遮蔽效應」廣義上指的是能量較大的聲音能夠掩蓋或抑制能量較小的聲音。在音頻信號處理中，遮蔽效應通常可以細分爲頻域遮蔽和時域遮蔽。頻域遮蔽指的是在頻率域中，能量較強頻帶聲音會掩蓋能量較弱的頻帶，而時域遮蔽則是指在時域中，可以分成兩種類型: 前遮蔽（pre-masking），主要是發生在能量較強的聲音出現之前的短時間內，能量較較弱的目標聲音會被遮蔽。以及後遮蔽（post-masking），發生在能量較強的聲音出現之後的一段時間內，較弱的目標聲音被遮蔽。對於食道語者而言，主要面臨的是時域遮蔽的問題。當食道語者使用食道發聲時，由於聲音的振幅和頻率特徵與正常語音存在差異，這會導致食道語中的聲音在時間域上出現遮蔽效應。具體而言，人類在接收到聲音信號後需要一定的時間進行處理，這意味著如果前後的聲音有較大的能量差異，則會導致前後聲音的辨識及處理上的困難。如 Figure 1 所示:

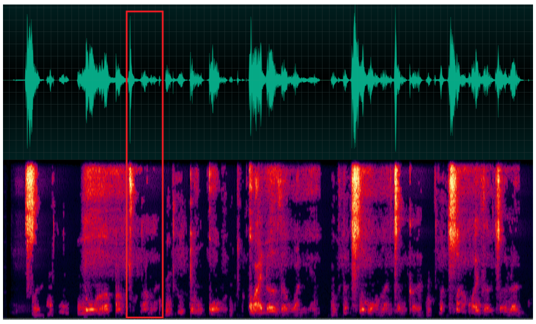


Figure 1: 食道語者語音的波形與頻譜

從 Figure 1 紅框部分可以發現當爆音出現後，較強的能量遮蔽了接在後面的母音導致可能出現了後遮蔽（post-masking），使得降低了整段話的可理解性。

基於上述觀察，本研究期望開發一款輕量語音演算方法，未來用於輕便的裝置，如領夾式麥克風，體積小巧可藏於衣領後方，通過過濾雜音並調整音量來放大被遮蔽的聲音和縮小爆音，從而提升食道語的音質和清晰度。這種聲音輔助裝置的應用有望可以改善食道語者的生活品質，使他們能夠更自信地表達自己的想法。

1.2 相關研究

綜觀食道語的研究，從早期食道語音增強技術基於線性預測編碼（Linear predictive coding, LPC）(Kenji et al., 2002) 和共振峰合成的傳統方法開始，通過對 30 毫秒的語音樣本，每 10 毫秒位移進行一次 LPC 和音高分析，窗口之間會有 20 毫秒的重疊，以提高食道語者的語音品質。儘管這些方法成功改善了音高的準確性，但由於食道語者的聲音特性不穩定，容易出現爆音問題，導致音高提取的穩定性受到影響，最終造成合成語音的品質下降。此外，此技術還帶來了較高的延遲，並且處理後的聲音與原始聲音有明顯差異，這些都限制了其在實際應用中的普及性和實用性。隨著技術的進步，研究逐漸轉向了基於統計模型的語音轉換方法，特別是高斯混合模型（Gaussian mixture model, GMM）的應用 (Doi et al., 2010)。GMM 技術通過對頻譜特徵、基頻和非週期訊號進行建模，並利用平行語料進行訓練，實現語音的轉換。

後來的研究進一步引入了多重回歸高斯混合模型（MR-GMM）和核回歸高斯混合模型（KR-GMM），並加入了語音品質控制向量，旨在提高食道語音的自然度和語音內容理解度 (Yamamoto et al., 2012)。這些方法通過控制語音的 male/female(gender), husky/clear (clearness), elder/younger(age),deep/thin(deepness) 等因素，成功提升了語音的個人化和適應性，但也對數據需求和計算資源提出了更高的要求。隨著深度學習技術的引入，研究者開始探索使用聲學隱性表徵（phonetic posteriorgrams, PPGs）(Chen et al., 2020) 結合各類模型來進行比較和食道語音的轉換。

總體而言，食道語音增強技術從傳統方法到統計模型，再到近年來的深度學習模型，顯示了技術發展帶來的顯著進步。然而，隨著技術的進步，這些方法的計算複雜度也逐漸增加。早期的 LPC 和共振峰合成方法相對計算簡單，但其缺點是容易受限於語音特性不穩定性，導致增強效果有限。隨後的 GMM 和 MR-GMM 方法通過更精細的建模，提高了語音的自然度和可懂性，但也導致了計算需求的增加，尤其是在多維特徵和大規模數據的訓練過程中。最近的深度學習技術，如 PPGs 結合各類神經網路模型，雖然在語音增強上展現了卓越的性能，但其龐大的計算量和對高效能硬體的依賴，成爲了在實際應用中推廣的瓶頸。

2 研究方法

2.1 系統目標及流程

我們的系統目標是透過聲學特徵提取和模型學習技術，專注於解決食道語者在發音過程中常遇到的「容易爆音的無聲子音」問題。這些無聲子音在語音中容易產生過大的振幅或過度的雜訊，從而影響語音的清晰度和可理解性。爲了有效地解決這些問題，系統將運用神經網路（neural network, NN）模型進行語音信號的分類和增強。模型將通過標註數據的訓練，學會辨識無聲子音中的爆音。系統主要目標是專注於降低語音訊號中過大的振幅，並有效地抑制食道語者吸氣所造成的吞口水雜訊，從而提升語音信號的整體音質，以確保語音在各種環境下都能保持清晰和自然。

2.2 系統架構

如 Figure 2 所示，系統主要分爲兩個階段：訓練階段和測試階段。在這個架構中，我們首先處理音檔數據，然後將其輸入到深度學習模型中進行分類。以下是系統架構的詳細說明：

訓練階段我們先對音檔做預處理，將 48kHz 取樣率的音檔降取樣至 16kHz 來進行本次實驗。並透過 Praat 軟體對音檔的每個段落進行標記，分配四種類型的標籤：1) 爆音（B）: 容易爆音的無聲子音、2) 正常語音（V）: 包含母音及其他子音、3) 雜音（N）和 4) 靜音（S）來讓模型學習特徵。並且抽取音檔的聲學特徵轉換成梅爾時頻譜 (Mel spectrogram)，其參數設置爲: 梅爾時頻譜的維度爲 80，音框長度爲 512 點，音框位移爲 160 點。我們將音檔每個音框以 50% 的重疊形式進行分割。並從音檔中提取出每 20 個音框 (約 0.032*20=0.64 秒) 組合成一張 20 × 80 維的 Mel spectrogram。這些 Mel spectrogram 與對應的標籤對齊後，輸入到神經網路（NN）中進行訓練。模型的輸出是四類分類：爆音（B）、正常語音（V）、雜音（N）和靜音（S）。

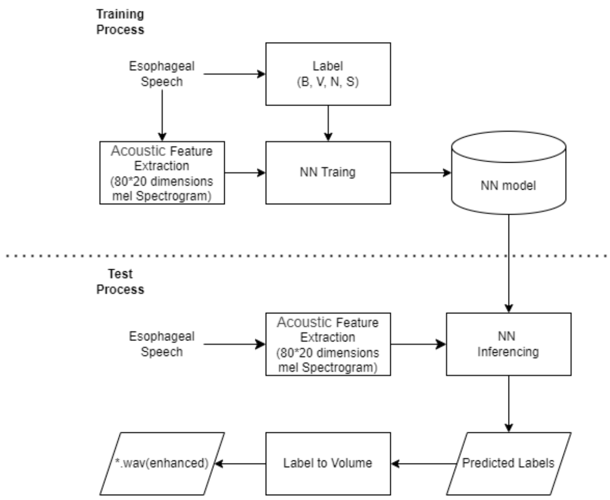


Figure 2: 系統架構圖

在測試階段，使用與訓練階段相同的數據預處理步驟來準備測試數據。然後將此輸入到已訓練好的模型中，獲取預測標籤（Predicted Labels）和加權音量（Volume）。爲了保證音檔的自然性，在分段處理時加入 Hann 窗以減少邊界效應，並重疊 50% 的形式，其數學式表示:

$$x_m[n] = x[n] \cdot \omega[n - mR] \quad (1)$$

其中，m 是當前窗的位置索引，$R=256$。$\omega(n)$ 代表 hanning window，數學式：

$$\omega[n] = \begin{cases} 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right) & \text{if } n \in [0, N-1] \\ 0 & \text{o.w.} \end{cases} \quad (2)$$

其中，$N=512$。Hann 窗有助於降低信號邊緣的強度，避免因分段造成的不連續性和微小爆音，保持信號的平滑性和自然性。在本研究中，我們對標籤進行了加權處理，以更好地反映不同類別的重要性。爲了實現這一點，我們爲每個標籤分配了相應的加權係數。其中，N

（噪音）和 S（靜音）被賦予加權係數爲 0，而 B（爆音）和 V（正常語音）分別被賦予加權係數爲 0.3 和 1。因爲我們希望可以抑制會爆音的無聲子音，來達到減少遮蔽效應的問題，其數學式表示:

$$y[n] = \sum_{m=0}^{M-1} g_m x_m[n] \tag{3}$$

其中 M 是總窗口數，g_m 是第 m 個窗口的加權係數，數學式表示:

$$g_m = \begin{cases} 1 & \text{if label}_m = \text{V} \\ 0.3 & \text{if label}_m = \text{B} \\ 0 & \text{if label}_m = \text{N or label}_m = \text{S} \end{cases} \tag{4}$$

2.3 模型訓練

2.3.1 CNN 架構

本研究所使用的卷積神經網路（CNN）架構設計如 Table 2 所示：模型接受形狀爲 20×80×1 的梅爾頻譜圖像作爲輸入，並通過兩個卷積層進行特徵提取。第一個 2 維卷積層使用 8 個 5×5 的卷積核，第二個 2 維卷積層使用 10 個 5×5 的卷積核，並使用 ReLU 激活函數。每個卷積層後面接有 2×2 的最大池化層來進行下採樣，並在每個卷積層後應用 60% 的 Dropout 操作以減少過度擬合。在經過展平層（Flatten）後，輸入到一個包含 128 個神經元的全連接層，該層也應用 Dropout。最終的輸出層包含 4 個神經元，使用 Softmax 激活函數進行多類別分類。使用 Adam 來做爲 Optimizers，學習率設爲 0.0001，損失函數選擇分類交叉熵（categorical crossentropy），並通過準確率（accuracy）來評估性能。

Layer	Output Shape	Param#
Conv2D	(20, 80, 8)	208
MaxPolling2D	(10, 40, 8)	0
Dropout	(10, 40, 10)	0
Conv2D	(5, 40, 10)	2010
MaxPolling2D	(5, 20, 10)	0
Dropout	(5, 20, 10)	0
Flatten	(1000)	0
Dense	(128)	128128
Dropout	(128)	0
Dense	(4)	516

Table 2: CNN 模型架構

2.3.2 GRU 架構

本研究也使用門控循環單元（GRU）架構進行對比設計如 Table 3 所示：模型將輸入的形狀從 (20, 80, 1) 重新塑形爲 (20, 80)，以適應後續的處理步驟。隨後，應用了一層一維卷積層（Conv1D），此層使用 128 個濾波器，核大小爲 3。接著，模型包含了三層 GRU 層，均設置爲 64 個單元，並加入了 60% 的 Dropout 層，以防止過度擬合。此後，輸出經過展平層（Flatten），並傳遞至一個包含 128 個神經元的全連接層，再次加入 60% 的 Dropout 層。最後，通過一個 Softmax 激活函數的輸出層，將輸入數據分類爲四個類別。爲了最佳化模型，我們選擇了 Adam 來做爲 Optimizers，設置學習率爲 0.00001。損失函數則採用了多類別交叉熵損失（categorical crossentropy），並將模型的評估指標設定爲準確率。

Layer	Output Shape	Param#
Reshape	(20, 80)	0
Conv1D	(20, 128)	30848
GRU	(20, 64)	37248
Dropout	(20, 64)	0
GRU	(20, 64)	24960
Dropout	(20, 64)	0
GRU	(64)	24960
Dropout	(64)	0
Flatten	(64)	0
Dense	(128)	8320
Dropout	(128)	0
Dense	(4)	516

Table 3: GRU 模型架構

3 實驗結果及討論

3.1 模型評估方法

爲了評估建構的模型性能，本研究我們採用了混淆矩陣（Confusion Matrix）作爲主要的評估工具並且計算了準確率（Accuracy）。混淆矩陣是一種常用的分類模型評估方法，拿預估結果（Predicted）和答案（True）進行對比，並計算出每個 label 各自的轉換率，這樣可以提供關於模型預測準確性和錯誤類型的詳細資訊。我們是多類別的分類模型，所以 Accuracy 爲:

$$\text{Accuracy} = \frac{\text{正確分類的樣本}}{\text{全部樣本}} \tag{5}$$

我們也計算了「個別」類別的 precision，數學式如下:

$$\text{Precision of class A} = \frac{\text{TP}_{\text{classA}}}{\text{TP}_{\text{classA}} + \text{FP}_{\text{classA}}} \tag{6}$$

3.2 實驗結果

實驗結果我們採取主觀層面跟客觀層面的分析。在客觀層面，是測試由食道語者提供的 105 句語音音檔，總計 18 分鐘左右，並使用我們訓練的模型輔助切割後並使用 Praat 軟體進行調整來得到各類別的切割位置，並將資料拆成 train 85 句、dev 10 句、test 10 句三個集合分辨用於訓練、評估、測試模型。在研究中，爲了證明 VoiceBank-2023 語料庫在語句數量較少的情況下，通過有順序地累積語句，可以更迅速且有效地涵蓋中文中的聲母和韻母，我們通過實驗比較了隨機選取和有序選取的訓練狀況。資料隨機選取的詳細資料如 Table 4 所示。

Sets	Duration	#Sentences	#Syllables
train	14	85	6393
dev	2	10	913
test	2	10	916

Table 4: 隨機選取資料集，食道語者各資料集音檔長度（分鐘）、句數、音節數

此外，按順序選取的資料集詳細資料如 Table 5 所示。

Sets	Duration	#Sentences	#Syllables
train	14	85	6654
dev	2	10	739
test	2	10	829

Table 5: 有序選取資料集，食道語者各資料集音檔長度（分鐘）、句數、音節數

在隨機選取的狀況下，使用 CNN 模型進行訓練後，我們獲得了如下的最佳結果如 Table 6 所示。

P\T	B	V	S	N	#labels
B	0.872	0.119	0.001	0.007	838
V	0.001	0.932	0.031	0.058	5625
S	0.001	0.024	0.915	0.058	2375
N	0.003	0.062	0.134	0.800	1866
Acc.	0.9005				

Table 6: Confusion matrix for CNN model trained with random-ordered corpus.

此外，我們也測試了 GRU 模型作爲對照，以評估不同模型架構在相同語料配置下的性能表現。經過相同的隨機選取和訓練過程，使用 GRU 模型的最佳結果如 Table 7 所示。

P\T	B	V	S	N	#labels
B	0.896	0.095	0.000	0.001	838
V	0.017	0.952	0.018	0.012	5625
S	0.001	0.049	0.858	0.092	2375
N	0.004	0.012	0.007	0.800	1866
Acc.	0.9003				

Table 7: Confusion matrix for GRU model trained with random-ordered corpus.

在按序選取的狀況下，CNN 模型的最佳結果如 Table 8 所示。

P\T	B	V	S	N	#labels
B	0.920	0.069	0.000	0.013	578
V	0.023	0.941	0.017	0.019	5567
S	0.002	0.031	0.888	0.078	2122
N	0.002	0.043	0.070	0.883	1482
Acc.	0.9195				

Table 8: Confusion matrix for CNN model trained with ordered corpus.

GRU 模型在按序選取的狀況下的最佳結果如 Table 9 所示。

P\T	B	V	S	N	#labels
B	0.927	0.055	0.002	0.016	578
V	0.025	0.932	0.015	0.027	5567
S	0.001	0.033	0.879	0.086	2122
N	0.003	0.023	0.076	0.898	1482
Acc.	0.9150				

Table 9: Confusion matrix for GRU model trained with ordered corpus.

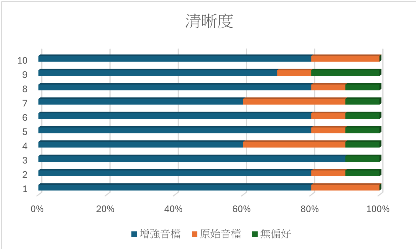
根據上述實驗結果，我們發現隨機選取語料會降低模型的準確度。因此，我們選擇了按照順序選取語料的策略，並對兩個模型的準確度和相關參數進行了詳細比較。以下是比較結果：

	CNN	GRU
B	0.920	**0.927**
V	**0.941**	0.932
S	**0.888**	0.879
N	0.883	**0.898**
Acc.	**0.9195**	0.9150
Network parameters	130k	127k

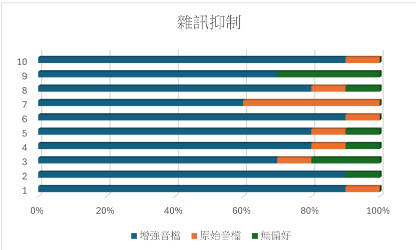
Table 10: Comparison of the CNN and GRU in the precisions of target labels, accuracies, and number of parameters.

在主觀層面的分析中，我們進行了聽感測

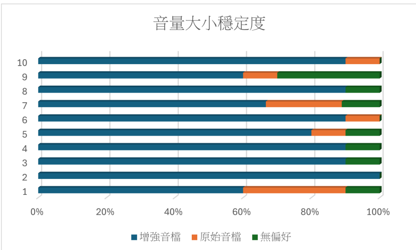
試。具體而言，我們將經過模型分類和語音增強處理後產生的音檔進行隨機排序，然後讓 10 位測試者試聽各 10 個原始音檔及增強後的音檔。測試者被要求在成對的音檔樣本中選擇一個他們認爲較優的樣本。如果測試者無法分辨兩個音檔之間的差異，他們可以選擇“無偏好”。我們根據三個方面進行詢問，進而得到 Figure 3 的結果。



(a) 清晰度



(b) 雜訊抑制



(c) 音量大小穩定

Figure 3: 三種評估指標分別是 (a) 清晰度、(b) 雜訊抑制、(c) 音量大小穩定

根據 Figure 3，我們可以觀察到大多數測試結果顯示，增強後的音檔在音量穩定性方面有顯著改善。然而，少數音檔在清晰度和雜訊抑制上表現未達預期，特別是某些音檔仍然存在雜訊或聲音斷斷續續的情況。儘管如此，整體結果仍顯示出正向的改善。

3.3 實驗討論

從實驗結果可以觀察到，使用 VoiceBank-2023 語料庫時，有序選取語料對模型準確度的提升非常顯著。

對於有序選取語料的情況，CNN 模型在正常語音（V）的準確率相對較高。部分辨認失敗的原因在於，雖然被歸類在正常語音（V）中的無聲子音不常出現爆音，但在極少數情況下仍有可能出現爆音。此外，可能由於類似遮蔽效應的存在，母音有時會被子音遮蔽，這可能導致在頻譜轉換後，母音被誤辨認爲雜音或靜音。儘管模型在大多數情況下能正確辨認正常語音（V），但仍需進一步改進以提高在此類情況下的辨識精度。至於靜音與雜音的辨認率，CNN 模型均接近九成，誤辨率主要表現爲雜音被辨認爲靜音，或靜音被辨認爲雜音。然而，由於我們將雜音與靜音的加權係數設爲 0，因此這些誤辨率對最終結果並不會造成實質性影響，即使在某些情況下發生誤辨識，也不會顯著影響整體性能。另一方面，GRU 模型在辨識爆音（B）方面有明顯更好的效果，這是因爲 GRU 可以利用其記憶單元捕捉前後文的信息，從而更好地理解爆音（B）在不同語境中的表現。這對於提高子音分類的準確度非常重要，也是 GRU 在子音辨識結果上表現良好的原因。相比之下，CNN 通過卷積提取局部特徵，這些特徵可能不足以捕捉爆音（B）的快速變化和短時動態，導致效果不如正常語音（V）。

至於根據聽覺測試結果，我們可以看到語音增強處理在不同的指標上有著不同的效果。在清晰度方面，大多數測試音檔的結果是增強後的音檔被受測者認爲更清晰，這表明增強處理在改善清晰度方面總體上是有效的。然而，在某幾個音檔約有 20% 到 30% 的受測者在清晰度評估中選擇了原始音檔，這可能是因爲增強過程中把一些被認爲是雜訊的正常語音 (V) 給濾掉，使得音檔的聽感變得斷斷續續，從而影響了清晰度的感受。在雜訊抑制方面，雖然大部分音檔經過增強後的效果是正面的，但有一個音檔卻有 40% 的受測者選擇了原始版本，這說明這個特定音檔在增強過程中可能未能有效地抑制雜訊，反而在某種程度上保留了原始雜訊。至於音量大小的穩定性，大多數受測者更傾向於選擇增強後的版本，這表明在這一指標上，增強處理成功地平衡了音量，使得音檔的音量聽起來更穩定。

總體而言，從三個方向來看，增強處理的效果都是正面的。特別是在音量大小的穩定性方面，處理得非常成功，大多數受測者都選擇了增強後的音檔，顯示出增強技術在平衡音量方面的顯著優勢。雖然在清晰度上有少數音檔的處理效果不如預期，但整體上仍然能是正面的。雜訊抑制方面也顯示出良好的效果，但某

些音檔在這方面的表現略顯不足，提示我們在未來的改進中需要更加關注這一指標。總的來說，這些結果表明增強技術在大多數情況下能有效提升音質，尤其是在音量穩定性上，未來可以進一步優化處理過程以提高清晰度和雜訊抑制的效果。

4 結論

在本研究中，根據上述實驗結果，考量到相對參數下計算量跟的訓練跟準度，我們最後選擇使用卷積神經網路（CNN）來識別食道語中的各種聲音成分，包括爆音、正常語音、靜音和雜音。實驗結果顯示，該模型在各類聲音成分的識別上均表現出色，在會爆音 (B) 跟正常語音 (V) 識別方面達到了高準確率。正常語音 (V) 的準確率相對較高，部分辨認失敗的原因在於有些無聲子音偶爾出現的爆音以及遮蔽效應的影響，導致母音有時被誤認爲雜音或靜音。靜音與雜音的辨認率接近九成，誤辨主要集中在雜音和靜音之間的相互誤認，這對最終結果影響不大。爆音的準確率達到了 92%，顯示了該模型在處理子音方面的強大能力。總體來說，該系統在辨認爆音、正常語音、靜音及雜音方面均表現良好，低計算量跟準確度也證明了其可行性和可靠性。未來的工作可以集中在進一步優化模型，特別是針對遮蔽效應和少數爆音情況進行改進，以提升系統的整體性能和準確性。此外，擴充語者和語句的多樣性也是一個重要方向，這樣可以提高模型在不同情境下的穩健性和適應性。通过這些措施，可以進一步提高語音增強技術的有效性和廣泛適用性。

這項研究除了開發在小型的發聲輔助裝置外，或許可以應用在一般麥克風上來消除雜音與爆音，甚至部分還可以結合手機或網頁開發成一個應用軟體，使人與人在用該軟體交流時可以避免突然有雜訊或爆音的情況發生。

References

Chen-Yu Chen, Wei-Zhong Zheng, Syu-Siang Wang, Yu Tsao, Pei-Chun Li, and Ying-Hui Lai. 2020. Enhancing intelligibility of dysarthric speech using gated convolutional-based voice conversion system. In *Interspeech*, pages 4686–4690.

Hironori Doi, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. 2010. Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models. *IEICE TRANSACTIONS on Information and Systems*, 93(9):2472–2482.

Kadria Ezzine, Joseph Di Martino, and Mondher Frikha. 2022. Intelligibility improvement of esophageal speech using sequence-to-sequence voice conversion with auditory attention. *Applied Sciences*, 12(14):7062.

Rachel Kaye, Christopher G Tang, and Catherine F Sinclair. 2017. The electrolarynx: voice restoration after total laryngectomy. *Medical devices: evidence and research*, pages 133–140.

Matsui Kenji, Hara Noriyo, Kobayashi Noriko, and Hirose Hajime. 2002. Enhancement of esophageal speech using formant synthesis. *Acoustical Science and Technology*, 23(2):69–76.

Luis Serrano, Sneha Raman, David Tavarez, Eva Navas, and Inma Hernaez. 2019. Parallel vs. non-parallel voice conversion for esophageal speech. In *INTERSPEECH*, pages 4549–4553.

Jia-Jyu Su, Pang-Chen Liao, Yen-Ting Lin, Wu-Hao Li, Guan-Ting Liou, Cheng-Che Kao, Wei-Cheng Chen, Jen-Chieh Chiang, Wen-Yang Chang, Pin-Han Lin, et al. 2023. Voicebank-2023: A multi-speaker mandarin speech corpus for constructing personalized tts systems for the speech impaired. In *2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Kenzo Yamamoto, Tomoki Toda, Hironori Doi, Hiroshi Saruwatari, and Kiyohiro Shikano. 2012. Statistical approach to voice quality control in esophageal speech enhancement. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4497–4500. IEEE.

運用長句簡化及少樣本學習以提升大型語言模型辨識蛋白質交互作用準確性

Enhancing Protein-Protein Interaction Recognition Accuracy in Large Language Models through Sentence Reduction and Few-Shot Learning

Yi-Yun Chou, Li-Kai Chen, Wei-Ren Liu, Hong-Jie Dai[*] and Ching-Tai Chen

Department of Bioinformatics and Medical Engineering, Asia University
[*]Department of Electrical Engineering, National Kaohsiung University of Science and Technology
choucandy998@gmail.com, a0920324955@gmail.com, sssss1050287@gmail.com,
hjdai@nkust.edu.tw, ctchen@asia.edu.tw

摘要

蛋白質交互作用（Protein-Protein Interaction，PPI）辨識是生醫文獻探勘的重要課題。近年來具備通用知識的預訓練大型語言模型（Large Language Model，LLM）能夠透過 Prompt 快速運用於下游任務的特性，已廣泛運用於各種自然語言處理議題。本研究透過 Prompt Engineering 提出 Sentence Reduction 和 Few-Shot 方法引導五種 LLM 進行 PPI 辨識，並驗證其效能，結果顯示本文的方法可有效提升 LLM 的效能，在三組語料上的 F1-Score 均達 0.8 以上，優於過往傳統方法，並揭示不同的 Prompt 策略和 Few-Shot 對提高準確度的重要影響。

Abstract

Protein-Protein Interaction (PPI) recognition is a crucial task in biomedical literature mining. In recent years, pre-trained Large Language Models (LLMs) with general knowledge have been widely applied to various natural language processing tasks due to their ability to quickly adapt to downstream tasks through prompting. This study proposes Sentence Reduction and Few-Shot methods through Prompt Engineering to guide five LLMs in PPI recognition and evaluates their performance. The results show that our approach can effectively enhance LLM performance, achieving F1-Scores above 0.8 on three datasets, outperforming traditional methods. The study also reveals the significant impact of different prompting strategies and few-shot learning on improving accuracy.

關鍵字：大型語言模型、文獻探勘、提詞工程、句子簡化、少樣本學習

Keywords: Large Language Model, Text Mining, Prompt Engineering, Sentence Reduction, Few-Shot Learning

1 Introduction

今日生物醫學研究領域中書目資料的數量持續大幅成長，這些文獻蘊含大量研究成果，對生醫領域具有重要參考價值（N.-W. Chang 2020），然而，面對浩如江海的資料量，研究人員往往需投入大量人力及時間以取得重要資訊，近年運用自然語言處理技術（Natural Language Processing，NLP）作生物醫學文獻探勘有重大貢獻，例如從文獻中自動提取生物實體間相互作用（Warikoo, Chang, and Ma 2022），其中蛋白質交互作用（Protein-Protein Interaction，PPI）的辨識，可為生醫研究、新藥開發、癌症免疫、精準治療等提供相當重要的資訊。目前研究的熱點為結合注意力機制和特徵向量並進行深度學習（Liu and Guo 2019），但這種這種方法需要調整大量的超參數，過程中不僅耗時，且重複訓練往往需要大量的計算資源，增加使用難度（Li et al. 2021）。而近年來大型語言模型（Large Language Model，LLM）的興起（Minaee et al. 2024），不但結合深度學習的優勢（Du et al. 2024; Wolf et al. 2020），同時可以藉由透過提詞工程（Prompt Engineering）設計並優化 Prompt（Schulhoff et al. 2024）提升 LLM 的學習能力（Chen et al. 2023），並進行各種下游任務，現今已在 NLP 領域上廣泛利用（Deng et al. 2022）。

本研究透過 Prompt Engineering，測試了三種不同 Prompt 的設計概念，分別為：（1） Sentence Reduction：主要包含刪除句子冗餘成分、合併簡化後句子、句法結構轉換、同義詞替換、以更精煉的描述取代原有敘述，重構句子成分（Jing and McKeown 2000）。透過從複雜文本中提煉出關鍵信息，將長句濃縮為精簡短句，同時保持原意（Nguyen et al. 2004; Feng et al. 2023）。本研究中，我們運用 LLM 實現 Sentence Reduction 並保留與 PPI 相關資訊；（2） Few-Shot Learning：透過提供少量與任務相關的範例，協助 LLM 透過類比學習指定任務的回答邏輯（Brown et al. 2020），指導 LLM 能夠更精確地進行下游目標執行任務；（3） One-Stage 及 Two-Stage Prompting：基於前述兩種方法，One-Stage Prompting 策略於提示中整合了 Sentence Reduction 與辨識 PPI 這兩個核心任務，而 Two-Stage Prompting 策略則將這兩個任務分開獨立執行。結合上述三項方法，我們共設計出 15 種 Prompt，並使用五個廣為人知的公開商用 LLM 及三組資料集進行驗證及性能比較。

2 Related Work

2.1 LLM 在 PPI 辨識任務上的應用

有多項相關研究展示了 LLM 在生物醫學文本探勘中的潛力。Park 等人（2023） 對 LLM 應用在 PPI 辨識任務上進行全面的評估，研究中使用多個 LLM 在 STRING 數據庫進行 PPI 辨識。他們設計兩項子任務：一是要求 LLM 生成給定語句中的蛋白質列表（PPI Task1），二是由 LLM 判斷兩個蛋白質是否存在相互作用（PPI Task2）。在 PPI Task1 中，研究者觀察到 LLM 傾向於根據給定蛋白質的首字母生成相關蛋白質名稱，導致 LLM 對相似名稱蛋白質能準確預測，但對不相似名稱的蛋白質則表現較差。對於 PPI Task2，研究者透過建構的平衡數據集評估 LLM 判斷語句中提及的兩蛋白質是否有 PPI 的能力，根據實驗結果 LLM 在 PPI Task2 可達 0.5-0.984 的 F1-Score，反映了 LLM 在二元分類的潛力。

同樣探討 LLM 在 PPI 辨識任務中的應用（Rehana et al., n.d.），該研究不僅驗證 BERT 模型在 PPI 識別任務中的高效能（在 LLL 語料庫可達 0.868 的 F1-Score），也展示了 GPT 模型在此領域中的應用價值。儘管 GPT 模型並非專門為生物醫學領域分析而設計，GPT4 在該研究中仍展現出與 BERT 相當的性能，在 LLL 語料庫達到 0.864 的 F1-Score，證明 GPT 模型具備從非結構化文本中有效識別 PPI 的能力。

2.2 Sentence Reduction 改善 PPI 識別效能

Jonnalagadda 和 Gonzalez（2009）開發一個名為 bioSimplify 的句子簡化工具，該工具基於語法規則執行以下四個主要步驟：移除無關短語、替換基因名稱、替換名詞短語以及基於依存關係的句子分割。在評估中，他們結合 bioSimplify 與 PIE 系統（Kim et al. 2008）做為基準 PPI 辨識工具進行實驗。根據結果顯示其系統的召回率（Recall）提高了 8%，F1-Score 提高 3 個百分點，凸顯句子簡化確可提升 PPI 抽取系統的效能。

3 Method

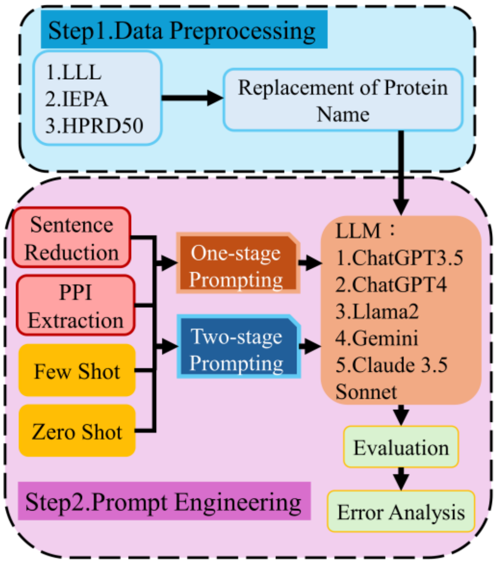


圖 1.實驗流程圖

儘管先前的研究已經證實 LLM 和 Sentence Reduction 在 PPI 識別任務上的有效性，但目前文獻中仍然缺乏將這兩種方法協同應用於提詞工程中以辨識 PPI。於此節我們闡述如何透過運用 LLM 實現 Sentence Reduction 並結合 Few-Shot Learning 指引 LLM 精確執行 PPI 辨識任務。圖 1 為本研究的實驗流程圖。在 Data Preprocessing 中，我們對 PPI 語料庫中的原始文本進行 Replacement of Protein Name，將蛋白質和基因名稱標記為特定的詞彙。在第二

個階段 Prompt Engineering 中，本研究以 Sentence Reduction 、PPI Extraction、Few-Shot Learning 為核心，開發出一系列不同的 Prompt，並把開發的 Prompt 分為一次作完 Sentence Reduction 及 PPI Extraction 的 One-Stage Prompting，與分別用兩次 Prompt 依序完成前述兩項任務的 Two-Stage Prompting。為驗證本文提出的 prompting 方法是否能優化 LLM 提取 PPI 的能力，搭配不同策略，本研究選用五個熱門的商用 LLM 進行測試，包含 Google Gemini Pro（Gemini Team et al. 2024），Meta Llama2 （Touvron et al. 2023, 2），Anthropic Claude3 Sonnet （Caruccio et al. 2024），OpenAI ChatGPT3.5，以及 OpenAI ChatGPT4。（Budzianowski and Vulić 2019; Floridi and Chiriatti 2020; Kalyan 2024; OpenAI et al. 2024; Wu et al. 2023），並對 LLM 的輸出進行評估與錯誤分析。以下章節針對前述流程中各子步驟進行詳細說明。

3.1 Dataset

本研究採用三個在 PPI 研究領域中廣泛使用的語料庫：LLL（Learning Language in Logic）（Nédellec and Nedellec, n.d.）、IEPA（Ding et al. 2002）、HPRD50（Peri et al. 2004）。這些語料庫被公認為評估 PPI 辨識準確率的標準數據集（Y.-C. Chang et al. 2016）。LLL 語料庫源於 2005 年同名研討會，其內容主要摘自 Medline 數據庫中的生物學文獻摘要，該語料庫的訓練集包含 269 個句子、測試集包含 61 個句子。IEPA 語料庫由 303 篇 PubMed 摘要組成，其中訓練集含有 681 個句子、測試集包含 136 個句子。HPRD50 語料庫的建構基於 Human Protein Reference Database，研究者從中隨機選取了 50 篇摘要，訓練集包含 363 個句子，測試集包含 70 個句子。

3.2 Replacement of Protein Name

在此步驟，我們根據語料庫中提供的蛋白質名稱標註資訊將每個句子中的任兩個蛋白質或基因名稱分別替換為 PROTEIN1 和 PROTEIN2 標籤。然而有些句子含超過兩個蛋白質或基因，對於其它蛋白質或基因名稱我們使用一系列泛化標籤，如 MOLECULE、MATERIAL、SUBSTANCE、PARTICLE、INGREDIENT 等進行替換，以保持句子語義結構，同時減少其餘非目標實體對 LLM 造成注意力分散的現象。對於包含多個蛋白質和基因名稱的句子，本研究採用排列組合的方法，這意味著同一原始句可能產生多個變體句子，每個變體句子中 PROTEIN1 和 PROTEIN2 標籤所代表的蛋白質實體會有所不同。透過這種方法能夠捕捉到同一語境下不同蛋白質對之間的潛在交互作用。在後續實驗中，這些通過排列組合的變體句子被視為獨立樣本（亦為 3.1 節中測試資料的樣本數），個別驗證正確率。

3.3 Prompt Engineering

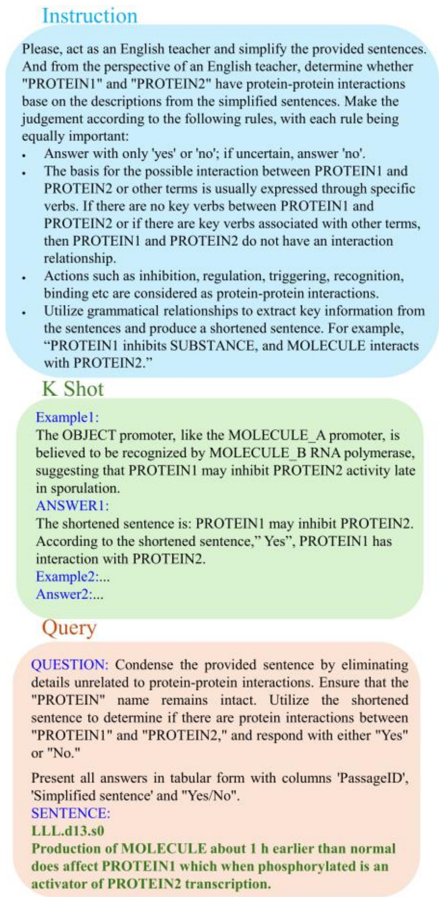


圖 2. Prompt 設計之模板

圖 2 顯示設計的 Prompt 主要核心模板，其包括兩個任務：第一個任務為 Sentence Reduction，執行句子簡化；第二個任務則是 PPI Extraction，專注於從文獻中識別並提取 PPI。我們將設計的 Prompt 模板分成三個部分，包含 Instruction 及 Few-shot Learning 以及最後一個部分的 Query：

1. Instruction：如圖 2 的 Instruction 部分所示，我們參考 Sivarajkumar 等人（2023）提出的 Heuristic Prompts 概

念，將任務轉化為一系列的指引規則，引導 LLM 完成任務。

2. Few-shot Learning：如圖 2 的 K Shot 部分所示，我們在此部分的模板中提供從 Training Data 抽取出來的 1~20 個固定範例做為 LLM 執行任務的參考。

3. Query：於此我們加入指令要求 LLM 以二元形式回答（Yes/No），並限制其輸出以表格方式呈現。最後於此部分提供要處理的句子給 LLM 執行任務。

3.4 Sentence Reduction

為了避免 LLM 被句子中過多的訊息分散注意力，本研究提出運用 Sentence Reduction 的方法讓 LLM 能夠基於語意及語法分析保留與蛋白質相關的句子片段，刪除其餘無關詞彙，透過這樣的方法來縮短冗長的句子，從而得到簡化的句子。為達成上述的目標，我們提出了三項規則：

- The essential information to be extracted pertains to the mutual interactions among PROTEIN1, PROTEIN2, MOLECULE, SUBSTANCE, ELEMENT, FACTOR, and MATERIAL. Information regarding interactions between any of the above items should be retained.

- Utilize grammatical relationships to extract key information from the Sentences and produce a shortened Sentence. For example, “PROTEIN1 inhibits SUBSTANCE, and MOLECULE interacts with PROTEIN2.”

- Actions such as inhibition, regulation, triggering, recognition, binding etc are considered as protein-protein interactions.

3.5 PPI Extraction

針對 PPI Extraction 的任務，我們提出以下規則：

- Answer with only 'yes' or 'no'; if uncertain, answer 'no'. Each sentence is independent; please make judgments only on the 'PROTEIN1' and 'PROTEIN2' from each sentence.

- If there is no direct interaction between "PROTEIN1" and "PROTEIN2" but they interact with MOLECULE, SUBSTANCE, ELEMENT, FACTOR, and MATERIAL, the answer is 'no,' as the focus is solely on the interaction between "PROTEIN1" and "PROTEIN2" in the sentence.

- The basis for the possible interaction between PROTEIN1 and PROTEIN2 or other terms is usually expressed through specific verbs. If there are no key verbs between PROTEIN1 and PROTEIN2 or if there are key verbs associated with other terms, then PROTEIN1 and PROTEIN2 do not have an interaction.

- Actions such as inhibition, regulation, triggering, recognition, binding etc are considered as protein-protein interactions.

3.6 One-Stage Prompting and Two-Stage Prompting

鑒於將複雜的任務分解成多個簡單的子任務可以優化 LLM 執行任務的能力（Schulhoff

Prompt	Method	Sentence Reduction	K shot for SR	PPI Extraction	K shot for PE
Prompt Type I	One-Stage Prompting	N	N	Y	K=0、1、3、10、20
Prompt Type II	One-Stage Prompting	Y	K=0、1、3、10、20	Y	K=0、1、3、10、20
Prompt Type III	Two-Stage Prompting	Y	K=3	Y	K=0、1、3、10、20

表 1.表格為實驗設計的 Prompt 實驗種類：Y 和 N 分別代表有或無運用此技術、K Shot 欄位中數字為 Few-Shot 的範例個數設置、Method 則為使用策略。SR 代表 Sentence Reduction，PE 代表 PPI Extraction et al. 2024），我們設計並實驗了兩種不同的策略：One-Stage Prompting 和 Two-Stage Prompting，這兩種策略的主要區別在於任務執行的結構。圖 2 所展示 Prompt 屬於 One-Stage Prompting，其將 Sentence Reduction 及 PPI Extraction 兩個任務整合為一個 Prompt 來進行任務，而 Two-Stage Prompting 則將圖 2 拆解成兩個 prompt；先進行 Sentence Reduction，從 LLM 得到簡化後的句子後，再透過第二個 Prompt 對簡化句進行 PPI Extraction。

3.7 Few-Shot Learning

在設計 Prompt 中，本研究採用 Few-Shot Learning，從 Training Data 中選取固定範例作為 K Shot，我們設置不同的 K 值以評估不同 Shot 數量對 LLM 學習效果的影響。表 1 展示了本研究設計的 Prompt 種類，Prompt Type I 為不含 Sentence Reduction 與 Few-Shot Learning 的 baseline 方法， Prompt Type II為 One-Stage Prompting， 在 Sentence Reduction 與 PPI Extraction 任務皆設置了 K=1、3、10、20 四種情況，使用上保持兩步驟的 K 值一致。Prompt Type III 為 Two-Stage Prompting，Sentence Reduction 的 K-Shot 設定為 K=3，因其相較於其它數字（1, 10, 20）準確率最高，PPI Extraction 的設置則與 Prompt Type I 相同。根據 Brown 等人（2020）提到 LLM 會因為 Prompt 中出現頻率較高的標籤而產生偏差，為了防止 LLM 在判斷時產生偏見，Prompt 設計特別考慮了範例中 Positive 和 Negative 例句的平衡，並制定了使用範例的比例規範：

- 當 K=1 時，提供一個 Positive 例子。
- 當 K=3 時，提供兩個 Positive 例子及一個 Negative 例子。
- 當 K=10 或 20 時，Positive 和 Negative 例子各占半數。

3.8 Zero-Shot Learning

除了前述的 K Shot 方法外，我們也嘗試了 Zero-Shot Learning；在這個方法中 LLM 僅接受描述任務的 Prompt（只包含圖 2 中的 Instruction 與 Query 的部分），而不提供任何範例。由於 Zero-Shot Learning 適用於缺乏大規模訓練數據的情境，僅向 LLM 提供任務的描述，此方法十分依賴 LLM 固有的能力（Radford et al., n.d.），因此在實驗中，我們藉由 Zero-Shot Learning 與結合 Sentence Reduction 及 Few-Shot Learning 的 Prompt 比較，以瞭解不同 LLM 對於 PPI 任務本身的理解程度，並探討 Few-Shot Learning 帶來的進步幅度。表 1 中的 Prompt Type I 設定 PPI Extraction 的 K=0 即為 Zero-Shot Learning。

3.9 Evaluation

本研究採用 Precision、Recall 及 F1-Score 指標評估所有實驗的結果。

$$Precision = \frac{the\ number\ of\ correctly\ recognized\ PPI}{the\ number\ of\ recognized\ PPI} \quad (1)$$

$$Recall = \frac{the\ number\ correctly\ recognized\ PPI}{the\ number\ of\ Actual\ PPI} \quad (2)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

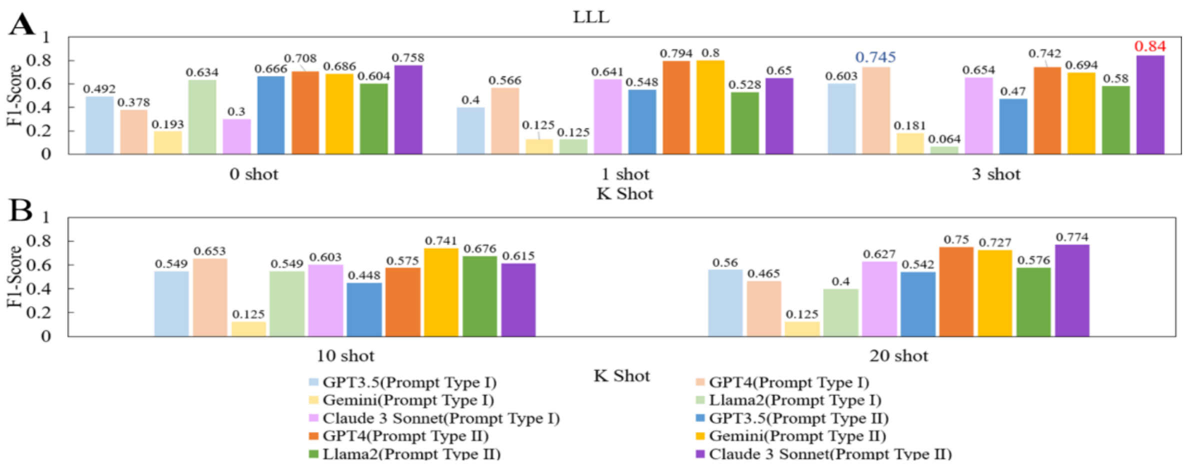


圖 3. 5 種 LLM 在 LLL 語料庫的測試結果，每個 X 軸刻度對應到 10 組 數據，左邊 5 組為 Prompt Type Ⅰ 實驗，右邊 5 組為 Prompt Type Ⅱ 實驗，F1-Score 在兩個實驗中的最高分分別以藍色、紅色文字標示，A 為 K=0、1、3 的測試結果，B 為 K=10、20 的測試結果

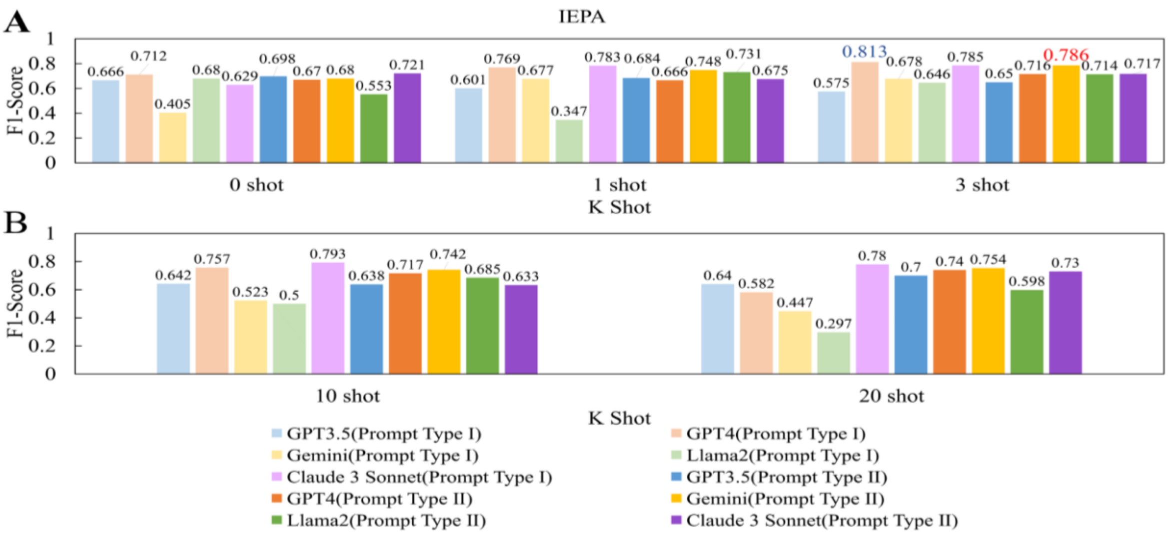


圖 4. 5 種 LLM 在 IEPA 語料庫的測試結果，每個 X 軸刻度對應到 10 組數據，左邊 5 組為 Prompt Type Ⅰ 實驗，右邊 5 組為 Prompt Type Ⅱ 實驗，F1-Score 在兩個實驗中的最高分分別以藍色、紅色文字標示，A 為 K=0、1、3 的測試結果，B 為 K=10、20 的測試結果

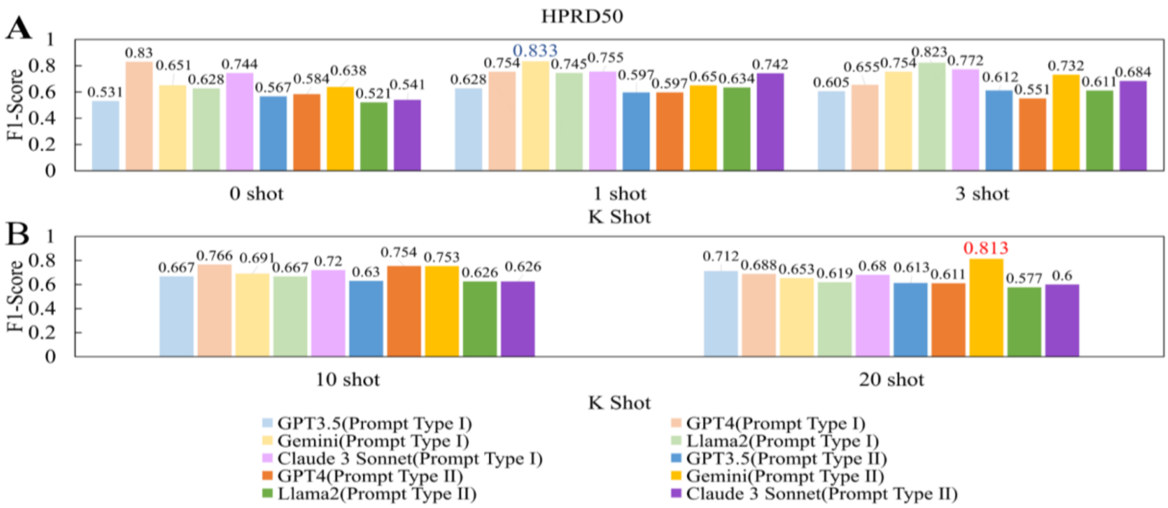


圖 5. 使用 5 種 LLM 在 HPRD50 語料庫的測試結果，每個 X 軸刻度對應到 10 組數據，左邊 5 組為 Prompt Type Ⅰ 實驗，右邊 5 組為 Prompt Type Ⅱ 實驗，F1-Score 在兩個實驗中的最高分分別以藍色、紅色文字標示，A 為 K=0、1、3 的測試結果，B 為 K=10、20 的測試結果

4 Result and Discussion

4.1 Sentence Reduction 對 LLM 識別 PPI 的優化程度

圖 3 到 5 為比較表 1 中 Prompt Type I（無 Sentence Reduction）和 Prompt Type II（應用 One-Stage Prompting 和 Sentence Reduction）分別在三個 PPI 語料庫上的實驗結果。LLL 語料庫的實驗結果如圖 3 所示，Prompt Type II 的 F1-Score 普遍高於 Prompt Type I，最高達 0.84，而 Prompt Type I 的結果大多低於 0.75，代表 Sentence Reduction 能顯著提升模型在 LLL 語料上的效能，提升幅度介於 0.01 至 0.5 之間。

圖 4 展示了 IEPA 語料庫的實驗結果。根據我們的統計，IEPA 是三個語料庫中語句平均長度最長的語料庫（LLL、IEPA 及 HPRD50 三個語料庫的測試集句子平均 token 數目分別為 34.31、35.4 與 25.73），令人意外的是，Prompt Type I 的 F1-Score 反而高於 Prompt Type II，最高甚至超過 0.8，顯示現今 LLM 在理解專業知識領域語句的卓越能力，然而 Prompt Type II 的結果在不同的 LLM 表現較為穩定，以 Llama2 為例，Prompt Type I 在 K=1 時 F1-Score 為 0.68，但在 K=3 時驟降至 0.064，波動極大。相比之下，Prompt Type II 的 Llama2 表現穩定維持在 0.55 以上，最高分與最低分差距僅 0.2，代表 Sentence Reduction 有

Dataset	LLL				IEPA				HPRD50			
Model	K*	P	R	F	K*	P	R	F	K*	P	R	F
GPT3.5	K=0	0.531	0.566	0.548	K=10	0.636	0.750	0.688	K=3	0.640	0.690	0.666
GPT4	K=10	0.727	0.533	0.640	K=1	0.633	0.803	0.708	K=10	0.720	0.692	0.705
Gemini	K=0	0.570	0.860	0.693	K=20	0.880	0.540	0.667	K=1	0.631	0.923	0.750
Llama2	K=10	0.750	0.300	0.429	K=0	0.500	0.642	0.562	K=3	0.471	0.961	0.632
C3S*	K=20	**0.880**	**0.733**	**0.800**	K=0	**0.910**	**0.750**	**0.823**	K=3	**0.790**	**0.730**	**0.760**

表 2.表格為使用 Prompt Type Ⅲ在 5 個 LLM 及 3 個語料庫上的實驗最好的結果，P、R、F 分別代表 Precision、Recall、F1-Score，表格中在 3 個語料庫上最高分以粗體標示。C3S*代表 Claude 3 Sonnet，K*代表最高分的實驗中 PPI Extraction 的 Few-Shot 設置

助於維持 LLM 在不同 K shot 之間的穩定度。圖 5 為 HPRD50 上的實驗結果，所有 LLM 在此語料庫上的表現均保持在 0.5 以上，且不似前兩個語料庫出現極大的分數落差。為了深入理解這一現象，我們對 HPRD50 語料庫中的句子結構進行分析，並發現其中的句子含有大量的括號，而透過 LLM 進行句子簡化時，往往會將括號刪除，改變了語法結構和語意內容，導致無法正確識別句中的 PPI。即使如此，在 Gemini 使用 Prompt Type II 仍有達到 F1-score 最高分 0.813，顯示 Sentence Reduction 結合 Few-Shot Learning 的效力。

綜合來說，結合 Sentence Reduction 技術的 Prompt 可讓 LLM 在 PPI 識別任務中表現出較高的穩定性。在大多情況下，分析簡化後的語能提升 LLM 識別 PPI 的準確性，此結果也顯示通過本研究的 Prompt Engineering 方式引導 LLM 執行 Sentence Reduction 是一種可行且有效的策略。

4.2 One-Stage Prompting 與 Two-Stage Prompting

表 2 呈現使用 Prompt Type III（應用 Two-Stage Prompting 和 Sentence Reduction）在 5 種 LLM 及 3 個語料庫上的最優結果。其中 Claude 3 Sonnet 在所有語料庫中均取得最高 F1-Scores：LLL 為 0.8，IEPA 為 0.823，HPRD50 為 0.76。對比圖 3 至 5 中 Prompt Type B 的結果可以發現 One-Stage Prompting 在 LLL 和 HPRD50 語料庫上的表現優於 Two-Stage Prompting。具體而言，One-Stage Prompting 在 LLL、IEPA 和 HPRD50 上分別達到最高 0.84、0.786 和 0.813 的 F1-Scores。綜合分析 Prompt Type II 和 C 的最高 F1-Scores 實驗數據表明 One-Stage Prompting 在 PPI 識別任務中整體上優於 Two-Stage Prompting。深入分析這一現象，我們發現兩者在 Sentence Reduction 步驟所產生的簡化句相近，然而，在 PPI Extraction 階段，Two-Stage Prompting 的錯誤率顯著高於 One-Stage Prompting。以 HPRD50 語料庫的實驗為例，在使用同樣 Claude 3 Sonnet 的情況下，針對句子

In contrast to PROTEIN1 PROTEIN2 did not interact with the AF2 domain of MOLECULE.

兩種 Prompting 策略均輸出簡化句

PROTEIN1 PROTEIN2 did not interact with MOLECULE.

但在 One-Stage Prompting 實驗中模型正確地判斷出 PROTEIN1 和 PROTEIN2 之間沒有相互作用，而在 Two-Stage Prompting 實驗中由於句子中出現了 "PROTEIN1"、"PROTEIN2" 及"interact with" 等關鍵詞，模型卻錯誤地認為 PROTEIN1 及 PROTEIN2 之間存在 PPI。這一結果表明 Prompting 策略的選擇對 LLM 在執行任務時的表現有些微影響，特別是在處理複雜度相對低的任務時，採用連續的、一體化的指令方法（如 One-Stage Prompting）似乎更能夠使 LLM 保持持續的注意力，從而提高預測的準確性。

4.3 Few-Shot 範例數量對效能的影響

由所有實驗結果來看，使用 Few-Shot 的方法普遍優於 Zero-Shot 的方法。以使用 Prompt Type II 的方法而言，對比所有 LLM 在 Zero-Shot 的 F1-score 最高分以及所有 LLM 採用 Few-Shot 的 F1-score 最高分，可以發現後者在 LLL、IEPA、與 HPRD50 語料庫的測試上分別提供 50% 、60%、及 70%的進步（圖 3、圖 4、圖 5），因此 Few-Shot Learning 可優化 LLM

Dataset	LLL			IEPA			HPRD50		
Method	P	R	F	P	R	F	P	R	F
LLM	**0.743**	**0.966**	**0.840**	**0.754**	**0.821**	**0.786**	**0.774**	**0.857**	**0.813**
SPBA	0.780	0.790	0.780	0.750	0.790	0.770	0.750	0.760	0.750
GK	0.725	0.872	0.768	0.696	0.827	0.751	0.643	0.658	0.634
LPTK	0.789	0.721	0.753	0.748	0.664	0.702	0.727	0.622	0.671

表 3. P、R、F 分別代表 Precision、Recall、F1-Score，LLM 採用 Prompt Type Ⅱ 方法的 LLM 最高 F1-score 作為代表，與其它三種非 LLM 的方法之效能比較。在各語料庫表現最高分的方法以粗體標記

識別 PPI 的準確度。但 Few-Shot 範例的數量增加不一定能優化 LLM 的表現，甚至有時會導致 LLM 的性能下降，如圖 4 中 Claude 3 Sonnet 的表現，在 K=0 時，F1-Score 來到了 0.758，但是在 K=1 時，F1-Score 卻下降到 0.65，而在 Shot 數目增加到 K=3 時，F1-Score 達到最高分數為 0.84。Few-shot Learning 並未隨著 K 值增加有所提升，我們推測可能提供的 K Shot 範例與 Query 中的句子並無類似的語法結構，導致 LLM 無法從中得到相關資訊，作正確的判斷邏輯。過去在 Shi 等人的研究中（2023），提到給予 LLM 與任務不相關的範例會分散 LLM 的注意力並大幅降低效能，與我們觀察的現象類似。而我們的實驗結果同時證實 prompt 內容對 LLM 造成輸出偏差的影響。在 K=1 的情境下，我們觀察到 LLM 輸出 False Positive 的數量偏高，而當 K=3，False Positive 才明顯降低，到 K=10、20，LLM 產生的 False Positive 與 False Negative 便趨於穩定。

4.4 與其它非 LLM 方法的效能比較

表 3 為我們使用 Prompt Type II 的方法與語法分析及統計準則方法進行比較，SPBA 為 N.-W. Chang （2020） 為建立準則樣本辨識文獻中 PPI 關係的方法；GK 為 Graph Kernel（Airola et al. 2008）根據句子的語法結構並建立及分析形式化的圖譜識別 PPI；LPTK（Warikoo, Chang, and Hsu 2018）為利用語義分析樹特徵提取 PPI。由表 3 可以看出我們的方法在 3 個語料庫上皆勝過其他方法，凸顯使用 Prompt Engineering 透過 LLM 識別 PPI 的方法在生醫文獻探勘上的巨大潛力。

5 Conclusion

本研究透過 Prompt Engineering 開發了適用於 PPI 的 Prompt，透過結合 Sentence Reduction 與 Few-Shot Learning 於開發的 Prompt 中來提升 PPI 的辨識效能。實驗證實我們所提出的 Sentence Reduction 方法在 LLL、IEPA、HPRD50 上最高的 F1-Score 分別可達 0.84、0.813 與 0.813，相較於沒有使用 Sentence Reduction 且 PPI Extraction 使用 Zero-Shot 的方法可分別提升 0.36-0.54、0.04-0.38 與 0.08-0.126 的 F1-Score，顯示本研究所提出的方法的有效性。另外，我們發現 Few-Shot 的準確率普遍較 Zero-Shot 更高，但當提供的範例量到達 10 甚至是更高的 20 筆並未有顯著的效能提升。此外，我們也比較 One-Stage 和 Two-Stage Prompting 兩種策略，結果顯示 One-Stage Prompting 策略通過整合任務流程，使 LLM 能更全面地把握任務本質，提高了注意力和準確性，準確率優於其它非 LLM 的傳統方法如語法分析及統計準則方法。綜合而論，我們的研究證實透過 Prompt Enginerring 配合公開可用之 LLM 進行 PPI 識別任務的可行性和巨大潛力，未來我們將探討 LLM 本身對於 PPI 任務的先天知識，並在生物醫學領域上使用 fine-tuning 方式優化 LLM，以及應用至其它生物醫學關係的抽取任務上，如：基因-疾病關聯、化學物質-蛋白質相互作用。

References

Airola, Antti, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. "All-Paths Graph Kernel for Protein-Protein Interaction Extraction with Evaluation of Cross-Corpus Learning." *BMC Bioinformatics* 9 （11）: S2. https://doi.org/10.1186/1471-2105-9-S11-S2.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." arXiv. http://arxiv.org/abs/2005.14165.

Budzianowski, Paweł, and Ivan Vulić. 2019. "Hello, It's GPT-2 -- How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems." arXiv. http://arxiv.org/abs/1907.05774.

Caruccio, Loredana, Stefano Cirillo, Giuseppe Polese, Giandomenico Solimando, Shanmugam Sundaramurthy, and Genoveffa Tortora. 2024. "Claude 2.0 Large Language Model: Tackling a Real-World Classification Problem with a New Iterative Prompt Engineering Approach." *Intelligent Systems with Applications* 21 （March）:200336. https://doi.org/10.1016/j.iswa.2024.200336.

Chang, Nai-Wen. 2020. "基於統計準則式方法偵測生醫文獻中的生物關聯." 臺灣大學生醫電子與資訊學研究所學位論文 2020 （January）:1–87. https://doi.org/10.6342/NTU202001123.

Chang, Yung-Chun, Chun-Han Chu, Yu-Chen Su, Chien Chin Chen, and Wen-Lian Hsu. 2016. "PIPE: A Protein–Protein Interaction Passage Extraction Module for BioCreative Challenge." *Database: The Journal of Biological Databases and Curation* 2016 （August）:baw101. https://doi.org/10.1093/database/baw101.

Chen, Banghao, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. "Unleashing the Potential of Prompt Engineering in Large Language Models: A Comprehensive Review." arXiv. https://doi.org/10.48550/arXiv.2310.14735.

Deng, Mingkai, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P. Xing, and Zhiting Hu. 2022. "RLPrompt: Optimizing Discrete Text Prompts with Reinforcement Learning." arXiv. http://arxiv.org/abs/2205.12548.

Ding, J., D. Berleant, D. Nettleton, and E. Wurtele. 2002. "Mining MEDLINE: Abstracts, Sentences, or Phrases?" *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 326–37. https://doi.org/10.1142/9789812799623_0031.

Du, Wenyu, Tongxu Luo, Zihan Qiu, Zeyu Huang, Yikang Shen, Reynold Cheng, Yike Guo, and Jie Fu. 2024. "Stacking Your Transformers: A Closer Look at Model Growth for Efficient LLM Pre-Training." arXiv. https://doi.org/10.48550/arXiv.2405.15319.

Feng, Yutao, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. "Sentence Simplification via Large Language Models." arXiv. http://arxiv.org/abs/2302.11957.

Floridi, Luciano, and Massimo Chiriatti. 2020. "GPT-3: Its Nature, Scope, Limits, and Consequences." *Minds and Machines* 30 （4）: 681–94. https://doi.org/10.1007/s11023-020-09548-1.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, et al. 2024. "Gemini: A Family of Highly Capable Multimodal Models." arXiv. http://arxiv.org/abs/2312.11805.

Jing, Hongyan, and Kathleen R. McKeown. 2000. "Cut and Paste Based Text Summarization." In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, 178–85. NAACL 2000. USA: Association for Computational Linguistics.

Jonnalagadda, Siddhartha, and Graciela Gonzalez. n.d. "Sentence Simplification Aids Protein-Protein Interaction Extraction."

Kalyan, Katikapalli Subramanyam. 2024. "A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4." *Natural Language Processing Journal* 6 （March）:100048. https://doi.org/10.1016/j.nlp.2023.100048.

Kim, Sun, Soo-Yong Shin, In-Hee Lee, Soo-Jin Kim, Ram Sriram, and Byoung-Tak Zhang. 2008. "PIE: An Online Prediction System for Protein-Protein Interactions from Text." *Nucleic Acids Research* 36 （Web Server issue）: W411-415. https://doi.org/10.1093/nar/gkn281.

Li, Yijing, Yanping Chen, Yongbin Qin, Ying Hu, Ruizhang Huang, and Qinghua Zheng. 2021. "Protein-Protein Interaction Relation Extraction Based on Multigranularity Semantic Fusion." *Journal of Biomedical Informatics* 123 （November）:103931. https://doi.org/10.1016/j.jbi.2021.103931.

Liu, Gang, and Jiabao Guo. 2019. "Bidirectional LSTM with Attention Mechanism and Convolutional Layer for Text Classification." *Neurocomputing* 337 （April）:325–38. https://doi.org/10.1016/j.neucom.2019.01.078.

Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. "Large Language Models: A Survey." arXiv. https://doi.org/10.48550/arXiv.2402.06196.

Nédellec, C, and Claire Nedellec. n.d. "Learning Language in Logic - Genic Interaction Extraction Challenge."

Nguyen, Minh Le, Susumu Horiguchi, Akira Shimazu, and Bao Tu Ho. 2004. "Example-Based Sentence Reduction Using the Hidden Markov Model." *ACM Transactions on Asian Language Information Processing* 3 （2）: 146–58. https://doi.org/10.1145/1034780.1034785.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2024. "GPT-4 Technical Report." arXiv. http://arxiv.org/abs/2303.08774.

Park, Gilchan, Byung-Jun Yoon, Xihaier Luo, Vanessa López-Marrero, Shinjae Yoo, and Shantenu Jha. 2023. "Comparative Performance Evaluation of Large Language Models for Extracting Molecular Interactions and Pathway Knowledge." arXiv. http://arxiv.org/abs/2307.08813.

Peri, Suraj, J. Daniel Navarro, Troels Z. Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, T. K. B. Gandhi, et al. 2004. "Human Protein Reference Database as a Discovery Resource for Proteomics." *Nucleic Acids Research* 32 （Database issue）: D497-501. https://doi.org/10.1093/nar/gkh070.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. n.d. "Language Models Are Unsupervised Multitask Learners."

Rehana, Hasin, Nur Bengisu Çam, Mert Basmaci, Jie Zheng, Christianah Jemiyo, Yongqun He, Arzucan Özgür, and Junguk Hur. n.d. "Evaluation of GPT and BERT-Based Models on Identifying Protein-Protein Interactions in Biomedical Text."

Schulhoff, Sander, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, et al. 2024. "The Prompt Report: A Systematic Survey of Prompting Techniques." arXiv. http://arxiv.org/abs/2406.06608.

Shi, Freda, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. "Large Language Models Can Be Easily Distracted by Irrelevant Context." arXiv. http://arxiv.org/abs/2302.00093.

0.1109/JAS.2023.123618.

Sivarajkumar, Sonish, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2023. "An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing." arXiv. https://doi.org/10.48550/arXiv.2309.08008.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. "Llama 2: Open Foundation and Fine-Tuned Chat Models." arXiv. http://arxiv.org/abs/2307.09288.

Warikoo, Neha, Yung-Chun Chang, and Wen-Lian Hsu. 2018. "LPTK: A Linguistic Pattern-Aware Dependency Tree Kernel Approach for the BioCreative VI CHEMPROT Task." *Database: The Journal of Biological Databases and Curation* 2018 （January）:bay108.
https://doi.org/10.1093/database/bay108.

Warikoo, Neha, Yung-Chun Chang, and Shang-Pin Ma. 2022. "Gradient Boosting over Linguistic-Pattern-Structured Trees for Learning Protein–Protein Interaction in the Biomedical Literature." *Applied Sciences* 12 （20）: 10199. https://doi.org/10.3390/app122010199.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. "HuggingFace's Transformers: State-of-the-Art Natural Language Processing." arXiv. http://arxiv.org/abs/1910.03771.

Wu, Tianyu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. "A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development." *IEEE/CAA Journal of Automatica Sinica* 10 （5）: 1122–36. https://doi.org/1

以商品名稱為基礎之跨電商平台的商品匹配
(Product Matching Across E-Commerce Platforms Based on Product Names)

李旭清 Xu-Qing Li; 鍾聖倫 Sheng-Luen Chung
Electrical Engineering Department
National Taiwan University of Science and Technology
Taipei, Taiwan
stanley890314@gmail.com; slchung@mail.ntust.edu.tw

摘要

本研究旨在解決消費者在不同電商平台上進行比價時所面臨的商品匹配問題。為此，本研究提出了一個兩階段網路架構：eComMatch，用於跨平台商品匹配。首先，通過三元組網路訓練的限縮 (Block) 網路進行初步過濾，然後由孿生網路訓練的匹配分類 (Match) 網路進一步確定兩件商品提及是否吻合。實驗結果顯示，透過我們所微調的語意編碼器 eComBERT，本兩階段模型除了能夠在合理的時間內匹配商品，還能在顧及計算量的考量下得到良好平衡的商品匹配精確性和效率，特別適用於需要快速更新和適應不斷變化市場的電商平台。

Abstract

This study aims to address the product matching issue faced by consumers when comparing prices across different e-commerce platforms. To solve this problem, this research proposes a two-stage framework: eComMatch, for cross-platform product matching. First, a blocking filtering process is conducted by a similarity network trained by triplet network, followed by further classification through a matching network trained with a Siamese network to determine if the product mentions match. Experimental results show that, with our fine-tuned semantic encoder, our two-stage model not only matches products within a reasonable time frame but also significantly improves the accuracy and efficiency of product matching. This method is particularly suitable for e-commerce platforms that require quick updates and adaptation to the ever-changing market.

關鍵字：電商平台、實體匹配、深度學習

Keywords: E-commerce platforms, Product matching, Deep learning

1 簡介

1.1 商品匹配

「商品匹配」(Product Matching) 指的是根據商品名稱，從一組字面上未必相同的商品名稱中，識別出與其實際上是相同商品的技術。在當今的電商環境中，商品匹配的重要性日益增加。原因在於，商家為了避免消費者比價，會透過更改商品名稱或描述呈現，以增加消費者在不同電家間搜尋同件商品的困難。舉例來說：在 PChome 上有一個商品「EPSON EF-11 雷射便攜投影機」，而該商品在 MOMO 上稱作「EPSON FullHD 雷射微型 3LCD 投影機 1000 流明」，如果是以字面上來看並無法直觀的認為兩者是相同的商品。

「單一比對」(Single Matching) 是商品匹配中最單純的任務：給定一個特定的商品，而要從一組數據庫中有效找出是否有匹配的實體。而涉及像是兩不同電商平台對應同一搜尋詞搜尋結果間多對多的商品配對 (mapping)，則是較普遍的商品匹配問題。

電商平台的消費者，會在跨電商平台之間尋找競品，藉此找到最優惠的價格，而電商營運業者則有需要透過商品匹配找到競品，藉此制定更靈活的商品定價策略，或是參考其他平台同件商品營銷的優缺點來改進自家平台。

1.2 貢獻

本研究提出了一個基於商品名稱的商品匹配方法，其為基於語意編碼器，採用限縮與分類的兩階段架構。該技術依序透過限縮 (Block) 網路和匹配分類 (Match) 網路進行商品匹配，其中限縮網路利用三元組網路訓練，負責將相似商品編碼群聚，而匹配分類網路則透過孿生網路訓練，負責界定出匹配商品。據此，本論文也提出如何由標註的匹配資料合成用來訓練各別網路所需的正負樣本的取樣策略。

作為說明示例，本論文聚焦於 B2C 情境下兩個競爭性電商平台之間的商品匹配問題。為此，本研究實現一匹配資料的標註界面，透過網路搜尋從兩競業電商平台取得的兩組匹配商品集，讓標註者能夠系統性地逐一標註與貯存匹配對組，然後按本論文提出訓練兩網路所需的正負樣本採樣策略，合成對應的子網路訓練資料。

1.3 本文架構

本論文以下各章節內容概述如下：第 2 節「商品匹配」包含了相關文獻的綜合回顧，其中包括實體匹配的研究、度量學習和深度學習在商品匹配中的應用。此外還討論了實體匹配與資料標註的關聯以及兩階段匹配方法的績效分析。第 3 節「兩階段的商品匹配」深入探討了 eComBERT 預訓練模型的訓練過程以及兩階段架構的設計，並介紹了正負樣本在訓練集中的取樣策略。隨後，第 4 節「實驗與結果」展示了如何在不同 B2C 平台之間測試商品匹配的效能，詳細描述了資料集的處理方式，並根據不同資料集和預訓練模型對兩階段模型進行訓練，展示了這些模型的效果比較。最後，第 5 節「結論」對研究成果進行了總結，並探討了其應用意義。

2 商品匹配

實體匹配：在實體匹配領域，許多研究致力於提高文本中的實體與知識庫中對應條目的匹配準確性。例如 Mihalcea and Csomai (2007) 利用維基百科作為實體連結的資源，提出了一個系統，該系統通過結合 tf-idf 和 Keyphraseness 方法自動提取關鍵字，並將文本中的重要概念自動連結到維基百科的相應頁面。這一系統的自動註釋效果良好，幾乎無法與人工註釋區分。然而，儘管該系統在註釋可靠性上取得了一定成功，但仍需進一步優化以處理更複雜的實體名稱變體和多義性問題。為了解決上述挑戰，Rao et al. (2013) 提出了使用線性回歸學習方法與豐富特徵集來進行實體比對，該技術不僅有效處理了實體名稱變體和多義性，還能預測何時不應進行匹配，顯著提升了實體匹配的準確性。然而，這些基於傳統機器學習方法的技術在處理結構化數據上的表現仍未達到最佳。在這一背景下，Mudgal et al. (2018) 回顧了許多文本處理中的相關匹配任務，並提出了包括 SIF、RNN、Attention 和 Hybrid 在內的四種解決方案。他們發現，儘管深度學習在處理結構化的實體匹配問題上並未超越當前的傳統解決方案，但在面對純文本或髒亂數據時，深度學習方法顯示出了顯著的優勢。為進一步推動實體匹配的研究，Wang et al. (2021) 則建立了一個通用的實體匹配基準，用以評估不同模型的效果，為後續研究提供了標準化的評測工具。

度量學習：實體匹配技術的不斷進步也帶動了度量學習的發展。Chopra et al. (2005) 提出了一種從數據中訓練相似度度量的方法。學習過程中，通過 Siamese 的網路最小化一個判別性損失函數來推動相似度度量，使得來自同一人的人臉對的相似度變小，來自不同人的人臉對的相似度變大。該網絡的架構旨在對幾何變形具有很強的魯棒性。在此基礎上，Hoffer and Ailon (2015) 引入了 Triplet 網路模型，該模型通過比較三個樣本的距離來學習有用的數據表示，並在不使用數據增強的情況下，顯著提高了分類準確率。此外，還有其他研究將 Siamese 架構與度量學習結合，Yuan et al. (2018) 提出了一個新的推薦系統框架，通過視覺特徵與類別特徵的融合，在匹配空間中學習物品之間的距離，從而更準確地識別匹配與不匹配的物品。Shah et al. (2018) 進一步利用 Siamese 神經網絡對 fastText 進行相似性訓練，從大量自行標註的數據中學習產品匹配，為產品匹配提供了一個新的視角，這也為後續的語意模型研究打下了基礎。

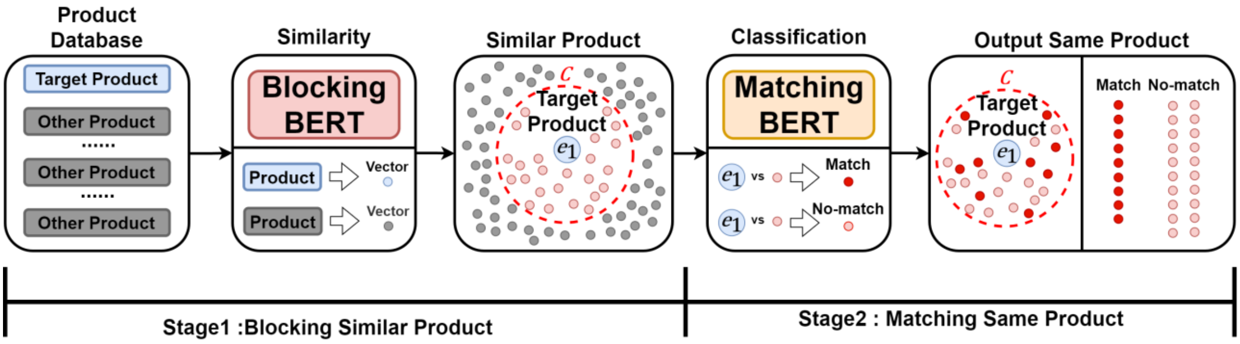


圖 1.兩階段匹配架構 eComMatch 與運作流程

語意模型：隨著度量學習技術的發展，語意模型在處理語意理解和匹配任務中的應用也越來越廣泛，並配合度量學習擁有更好的效能。最一開始，Vaswani (2017) 提出了一個基於 Attention 機制的 Transformer 語意模型，該模型能夠動態聚焦於輸入數據中最相關的部分，從而在語意理解方面展現出色的效果。這一成果為後續的 Transformer 模型奠定了基礎，BERT (Devlin, 2018) 應運而生，作為一種基於 Transformer 架構的自然語言處理模型，BERT 通過雙向編碼器設計同時考慮句子中每個詞的前後文語境，顯著提高了語意理解的精度。BERT 的強大性能主要來自於預訓練和微調兩個階段，這使得該模型在多種自然語言處理任務中表現出色。

進一步的研究中，SBERT (Reimers, 2019) 透過在 BERT 上添加 Siamese 網絡結構，使其能夠有效計算句子嵌入，並通過餘弦相似度快速計算大量句子之間的相似度，大幅提升了處理效率和效能。SBERT 在文本分類、語義搜索和對話系統中均表現卓越，並通過引入三元組和雙元組損失函數，進一步提升了模型的精度。Peeters and Bizer (2022) 進一步擴展了監督對比學習在產品匹配中的應用，並特別訓練了 BERT 模型。這顯示了對比學習在這一領域的潛力，同時也利用了多個數據集進行了測試。除此之外，Lin and Chen (2022) 提出了一種結合 ALBERT 的模型和 FastText 的模型的聯合方法 ALFA Matcher。該模型通過多頭自注意力機制篩選出關鍵信息一致的候選項，並利用 FastText 精確捕捉數據細節，在解決實體匹配任務上展現了優越的表現，進一步證明了語意模型在實體匹配中的重要性。

本論文做法：針對以商品名稱為匹配基礎的任務，我們提出了一個基於 Bert 文字編碼的兩階段的商品匹配架構：eComMatch，包括限縮 (blocking) 與匹配 (matching) 兩階段，如圖 1 所示。在 B2C 電商平台的商品匹配情境中，給定平台 A 中之一商品 e^1，我們要找出其在平台 B 中可能有多個的所有相對應的匹配商品 e^2，首先我們利用基於 Triplet Network 訓練的相似性 (Similarity) 比對網路，限縮以過濾掉 B 中較不相關的候選商品。隨後再使用基於 Siamese Network 訓練的匹配分類模型 (Matching Model) 進行匹配。針對以文字為主的商品名稱詞條，在設計以上相似性對比網路，以及匹配分類網路時，我們都採用特別電商情境文本訓練的 eComBert 作為 blocking 與 matching 這兩階段網路的預訓練模型，以有效縮小候選範圍，從而提升整體系統的性能。

，也就是先限縮再比對，兩階段比對流程如圖 1 所示。

2.1 實體匹配的定義

本研究討論 B2C 的電商情境中兩平台間的商品匹配問題。借用文獻中實體匹配的相關定義，並配合應用深度學習技術解決商品匹配時所需的資料標註與訓練、測試等的資料需求，整理出以下的符號。

符號：對應 B2C 的兩平台電商 A 與 B，我們用符號 D 與 D'表示來自各別對應平台的商品集合，我們用 $e_1 \in D$ 以及 $e_2 \in D'$ 來表示各別商品集合中的商品實體提及 (entity mention)。明確來說：B2C 的商品匹配任務是針對經過挑選某些規則挑選後，有可能匹配的候選商品集 (candidate set)：$C \subseteq D \times D'$ 中的各 $(e_1 \in D, e_2 \in D')$ 的組合，決定 e_1 與 e_2 是否 “match” 或是 “no − match”。

訓練資料：應對以上的分類問題，本研究採取「兩階段匹配架構 eComMatch」的深度學習方法解決。一般深度學習技術，配合訓練的需求，需要先有標註資料(tag)：$T = \{(e_i^1, e_i^2, l_i)\}_{(i=1)}^{|T|}$，其中的 l_i 即為二分類的 "match", "no − match"。在本研究中，我們只選擇性地標註正樣本，即 "match" 的資料：$\{(e_i^1, e_i^2, \text{match})\}_{(i=1)}^{|T|}$。在標註資料的呈現上，我們稱每對正樣本對 (positive matched sample pair) $(e_i^1, e_i^2, \text{"match"})$ 中的 e_i^1 為 Root、e_i^2 為 Leaf，分別來自 D 與 D'。

測試資料：另一方面，為了對訓練好的模型進行測試，我們需要另外一組與訓練資料完全不重疊的測試標註資料，我們之為 $T' = \{(e_i^3, e_i^4, l_i)\}_{(i=1)}^{|T'|}$，其中 $e_i^3 \in D''$，$e_i^4 \in D'''$；為了確保訓練集與測試集中的商品不重疊，我們另外要求：D'' 代表來自平台 A 的另一個產品集，且 $D \cap D'' = \emptyset$；同時 D''' 代表來自平台 B 的另一個產品集，且 $D' \cap D''' = \emptyset$。

檢索的定義：從搜尋的角度來看，如果將平台 A 上的商品視為查詢 (query)，而平台 B 上的商品視為被檢索的文件 (document)，則上述的：$D = D_{train}^q$；$D' = D_{train}^d$；$D'' = D_{test}^q$；$D''' = D_{test}^d$。其中 D 和 D'代表的是訓練時使用的商品集，而 D'' 和 D''' 代表的是測試模型時使用的商品集，在這裡 D 與 D' 和 D'' 與 D''' 皆為互斥關係 $(D \cup D') \cap (D'' \cup D''') = \emptyset$。正樣本對 $(e_i^1, e_i^2, \text{"match"})$ 中的 e_i^1 即為 query、而 e_i^2 即為 matched document，分別來自 D 與 D'。對於標註好正樣本對 e_i^1, e_i^2, "match" 中的 Root e_i^1 即為 query，而 Leaf 的 e_i^2 即為 matched document。

2.2 匹配績效指標

在本節中，我們將詳細說明商品比對兩階段模型的評估方式和所使用的評測指標。兩階段模型 eComMatch 包括第一階段的 Blocking 模型和第二階段的 Matching 模型。每個階段的評估方式各有側重，以下將具體介紹。

Blocking 模型的評估方式：在第一階段的 Blocking 模型中，我們的主要目標是評估模型在辨識實際相同商品時的效果。我們使用 B2C_Block_test 資料集進行測試，並在第 4.2 節中詳細說明該資料集的處理流程。在評估 Blocking 模型表現時，我們採用 Recall@K 作為主要指標，其具體計算公式如下：

$$\text{Recall@K} = \frac{|\{\text{Relevant items}\} \cap \{\text{TopK retrieved items}\}|}{\{\text{Relevant items}\}} \quad (1)$$

其中 {Relevant items} 代表所有實際相同的商品集合。{TopK retrieved items}代表前 K 個檢索結果的商品集合。這個指標幫助我們評估模型在初步過濾階段的效果，確保在大量的候選商品中能夠有效地檢索到實際相同的商品。

Matching 模型和兩階段的評估方式：在第二階段的 Matching 模型當中，我們使用 B2C_Match_Test 做為測試資料。針對 B2C 的 $D \rightarrow D'$ 多對多商品匹配問題，績效是針對所有 $D \times D'$ 組合，按判定 "$match$" 或是 "$no - match$" 分類結果與實際為匹配的 $(e^1, e^2, \text{"match"})$ 所計算出的 precision, recall, accuracy 以及 F1 score 決定。

3 兩階段的商品匹配架構

在電子商務平台中，如何快速且準確地找到匹配的商品是影響用戶體驗與交易成功率的重要因素。傳統的商品匹配方法通常是直接對比候選商品，這不僅耗時還可能導致匹配效果不理想。為了解決這一問題，我們提出了一種兩階段的商品匹配架構。在我們的架構中，我們首先採用了基於 Triplet Network 的限縮模型 (Blocking Model)，以過濾掉不相關的候選商品。隨後我們使用基於 Siamese Network 的匹配模型 (Matching Model) 來進行精細匹配。這樣的兩階段設計能夠有效縮小候選範圍，從而提升整體系統的性能。

3.1 eComBERT

目前已有多種基於不同語言和領域的 BERT 預訓練模型。，在台灣 CKIP Lab 中文詞知識庫小組有提供透過 ZhWiki 與 CNA 資料集進行預訓練的模型。然而，由於領域的差異 CKIP-BERT 並不適用於中文的電子商務環境。因此本論文使用 2 億 6 千多萬筆的中文電子商務領域的文本進行再次預訓練，最終提出了一個

適應於中文電子商務文本特徵的模型 eComBERT。

這樣的設計不僅繼承了 BERT 模型的優勢，還針對中文電子商務領域的特點進行了專門調整，使模型在這一特定領域的應用中能夠展現出更高的準確度和效能。透過這一過程，本研究展示了如何通過結合預訓練和轉移學習，針對特定領域需求進行模型優化，從而提升下游任務的表現

3.2 兩階段架構: eComMatch

本文針對商品匹配問題提出了一個兩階段的網路架構 eComMatch，解決如何從集合中找到與給定商品 e^1 匹配的商品 e^2。相比於傳統逐一比對集合中的所有商品，我們所設計的架構如圖 1 所示，透過先過濾後匹配的策略提升效率。第一階段利用經 Triplet 網路訓練的限縮模型 (Block) 進行初步篩選，過濾出可能匹配的候選商品。接著，第二階段使用由 Siamese 網路訓練的匹配模型進行更精確的匹配判斷，最終確定兩個商品描述是否一致。

在這兩種網路模型中，Siamese 和 Triplet 都屬於度量學習模型，專注於透過樣本的相似性訓練，提升模型對同類樣本間區別的敏銳度。其中 Siamese 網路每次輸入一對樣本，樣本可能屬於同類或異類，以學習樣本之間的相似度或差異性。而 Triplet 網路則同時輸入三個樣本：一個錨定樣本、一個正樣本和一個負樣本，透過優化三者之間的距離來進行訓練。本文詳細說明了兩階段架構中各模型的角色，其中，第一階段的 Triplet 模型負責排序和過濾，第二階段的 Siamese 模型負責分類與最終匹配判斷。

Siamese Network (Matching 模型)：Siamese 網路的架構設計會將兩個需要比較的樣本分別輸入到共享權重的子網路中，經過相同的特徵提取過程後生成對應的特徵向量。接著，這些特徵向量會被輸入到分類器中，並最終輸出 “match” 或 “no − match” 的分類結果。由於兩個子網路之間共享相同的參數權重，該結構能確保兩個樣本在同一特徵空間中進行比較，從而提高模型對樣本相似性的敏感度及判斷精確度。

在本文提出的兩階段架構中，第二階段的 Matching 網路模型使用 Binary Cross Entropy 作為損失函數來進行訓練。Binary Cross Entropy 損失函數是一種適用於二分類問題的標準損失函數，其目標是通過最小化模型預測值與真實標籤之間的差距來優化模型。該損失函數的數學定義如下：

$$Loss = -(ylog(\hat{y}) + (1-y)log(1-\hat{y})) \quad (2)$$

Triplet Network (Blocking 模型)：在第一階段的商品初步過濾中，我們採用了 Triplet 網路結構進行模型訓練。Triplet 網路的核心思想是利用三個元素 (Anchor、Positive、Negative) 組成的樣本組合進行訓練，以優化樣本間的相對距離。這種設計能夠有效提升模型對相似與不相似樣本的區分能力。

在 Blocking 模型中，我們使用 Triplet Loss 作為損失函數進行訓練。Triplet Loss 的主要目標是通過強化錨點與正負樣本之間的距離關係，使得模型能夠正確學習樣本之間的相似性和差異性。具體來說，Triplet Loss 希望最小化錨點與正樣本之間的距離，同時最大化錨點與負樣本之間的距離。其損失函數定義如下：

$$Loss(a,p,n) = max(0, d(a,p) - d(a,n) + \alpha) \quad (3)$$

3.3 訓練集正負樣本取樣策略

eComMatch 的模型訓練是透過個別訓練 Triplet 網路與 Siamese 網路完成，而個別網路的訓練資料是由反應是否匹配的正負樣本組合而成。正負樣本的來源為先前於提到的標註資料 $T = \{(e_i^1, e_i^2, l_i)\}_{i=1}^{|T|}$。正負樣本的取樣難度直接影響訓練效果：太簡單的樣本不能有效提升發路的鑑別度；而太難的樣本又容易造成網路的學習不穩定。適當的樣本才能夠顯著提升模型的準確度。

Triplet 的訓練樣本，對應之後第 4.1 節中的 B2C_block_Train，是由三元組資料 (Positive, Anchor, Negative) 組成，而 Siamese 的訓練樣本由(Anchor, Positive) 或 (Anchor, Negative) 組成：Siamese 的訓練資料則可以從 Triplet 的樣本中切割得來。以下說明如何收集到的資料製作成 (e^1, e^2,“match”) 樣式的標註正樣本集，透過兩種樣本取樣策略，建構 Triplet (Block) 和 Siamese (Match) 的訓練集。

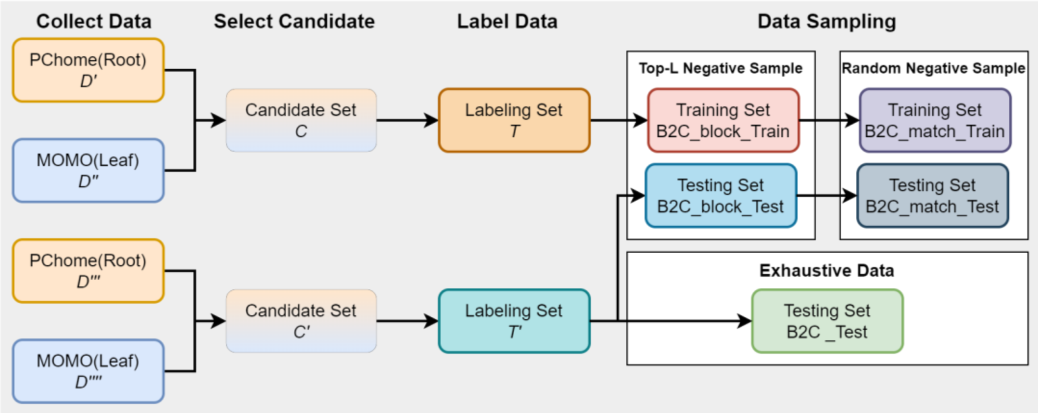


圖 2.訓練資料與測試資料處理流程

字面上最像的負樣本(Lexical Top-L Negative Sample)：首先我們透過標記先取出所有的正樣本組成(Anchor, Positive)，而在選擇負(Negative)樣本時，為了讓模型能學習認出字面上相似的負樣本商品名稱，我們採取 Lexical Top-L Negative Sample 策略，用意是為了從其他商品中利用輕量化的 BM25 算法找到字面上最相似，而實際上不同的商品作為負樣本。在這裡 Top-L 的 L 值取決於正樣本的數量。具體取樣步驟如下。

1. 詞距計算：同樣地對於每一對商品名稱，我們計算其 BM25 詞距分數。
2. 選取最相似樣本：根據計算出的 BM25 詞距，我們選取字面上最相似的前 L 筆樣本作為負樣本。
3. 樣本組合：這些負樣本將與 (Anchor, Positive) 樣本一起組成 L 筆的訓練樣本 (Anchor, Positive, Negative)，通過 Triplet Loss 來訓練模型。

簡易隨機負樣本(Simple Random Negative Sample)：用於 Siamese 分類模型的訓練資料，對應之後第四節中的 B2C_match_Train，可由先前組合成的 Triplet 的訓練樣本直接取得，並另外再使用本簡易隨機負樣本法。此方法旨在透過額外生成簡單負樣本，避免模型僅針對困難樣本進行訓練，從而提升模型在簡單情境 (例如完全不相關的商品描述) 下的泛化能力。具體來說，我們會基於已經生成的 N 筆 Triplet 訓練資料，透過以下步驟獲取兩倍於 N 的簡單負樣本 $2N$，詳細操作流程如下。

1. 正負樣本拆分：將 Triplet 的 N 筆訓練資料 $(\text{Anchor}, \text{Positive}, \text{Negative})$ 拆分成 N 筆 $(\text{Anchor}, \text{Positive}, \text{“match”})$ 和 N 筆 $(Anchor, Negative, \text{“no} - \text{match”})$，總共能拆分出 $2N$ 筆訓練資料。
2. 隨機選取：對於每個 Anchor 樣本，我們從其他商品中隨機選擇兩個樣本作為簡單負樣本，總共再取得 $2N$ 筆訓練資料。
3. 樣本組合：這些簡單負樣本將與 Anchor 樣本和正樣本一起組成訓練樣本，增加了 $2N$ 筆的簡單負樣本。

4 實驗與結果

為了進行實驗，我們首先需要明確設定實驗的情境，並建立相應的資料集。準備完成後我們才能對兩階段模型的效果進行評估。

4.1 情境：跨平台的商品配對

我們以 PChome 與 MOMO 這兩個 B2C 平台為例，說明如何解決兩平台間的商品配對問題，並檢視本技術測試商品配對的成效。我們總共有四個商品集合 D、D'、D'' 和 D'''，其中訓練的商品集為 $(D \cup D')$，測試的商品集為 $(D'' \cup D''')$，為確保訓練與測試的獨立性，訓練和測試商品集為互斥關係$(D \cup D') \cap (D'' \cup D''') = \emptyset$。符號 e 則用來表示商品的具體實體提及，其中 $(e^1 \in D, e^2 \in D', e^3 \in D'', e^4 \in D''')$ 意味著每一個實體提及都是各自商品集合中的元素。

接下來，我們使用獨立於訓練數據集的商品集合 D'' 和 D''' 來製作兩階段的評估測試集。首先從商品集合 D'' 中選擇一個商品 $e_i^3 \in D''$，並從 D''' 中找到所有的相同商品，構成測試數據集，其資料的處理流程如圖 2 所示。

4.2 資料集

eComMatch 是由 Triplet (Block) 與 Siamese (Match) 兩個網路串接而成。雖然它們個別獨立訓練和驗收，但最後是否符合電商商品匹配的任務，則需要合併這兩個網路以整體 eComMatch 來看。

我們首先從 PChome 和 MOMO 這兩個電商平台上蒐集並標註了 7,691 筆商品資料，其中 3,526 筆商品 ($D \cup D''$) 為完全不同的商品 (稱為 Root)，其餘 4,165 筆商品 ($D' \cup D'''$) 則是各自與 Root 商品相同的商品，我們將這部分稱為 Leaf (如 2.1 所定義)。在這裡我們透過 D 和 D'' 標註的資料為 $T = \{(e_i^1, e_i^2, \text{match})\}_{(i=1)}^{|T|}$，透過 D' 和 D''' 標註的資料則是 $T' = \{(e_i^3, e_i^4, l_i)\}_{(i=1)}^{|T'|}$。

為了構建資料集，我們將商品依照 Root 分成訓練集與測試集，比例為 5:1。訓練集僅包含 D 與 D'，測試集僅包含 D'' 與 D'''。在這裡，我們確保訓練與測試資料集互不重疊。基於本論文所處理的 B2C 任務場景，我們最終構建了五種資料集，詳細說明如下。

B2C_Block_Train：為了獲取 Block 模型的訓練資料，我們根據 3.3 節描述的負樣本取樣方法，透過 2,938 筆商品 D 從 3,329 筆商品 D' 提取了 4,088 筆三元組資料 (Negative, Anchor, Positive)，並將其納入 B2C_Block_Train 資料集中。

B2C_Block_Test：在處理完訓練集後，考慮到 Block 模型的主要目的是排序，我們直接將資料集 D'' 和 D''' 組合形成 B2C_Block_Test 資料集。在使用這些資料時，我們選擇了 588 筆的商品資料集 D'' 作為排序檢索的查詢數據，並將其與 836 筆商品資料集 D''' 進行排序。最後通過標註結果來評估 Recall@K 的分數。

B2C_ Match _Train：針對第二階段的 Match 模型我們也需要準備資料集，除了使用 3.3 節中提到的負樣本取樣方法外，我們還根據 3.3 節描述的簡單隨機負樣本方法額外取得訓練樣本。首先我們將 B2C_Block_Train 資料集拆解為適用於 Match 分類模型的樣本，生成了 4,088 筆 (Anchor, Positive, "match") 和 4,088 筆 (Anchor, Negative, "no-match") 的數據。接著，透過 3.3 的方法我們隨機抽取了 $2 * 4{,}088 = 8{,}176$ 筆負樣本，與前述的數據結合，最終構建了包含 16,352 筆訓練資料的 B2C_Match_Train 資料集中。所有用來訓練與測試 eComMatch 兩階段的資料都是由我們針對 PCHome 與 Momo 兩平台上爬文，然後經由人眼比對後所產的 4,165 筆的正樣本對 (positive matched sample pair) $(e_i^1, e_i^2, \text{"match"})$ 所組成的 $T = \{(e_i^1, e_i^2, l_i)\}_{(i=1)}^{|T|}$，再經由上節中的正負樣本取樣策略所組成。

B2C_ Match _Test：同樣地，我們可以從 D 和 D' 資料集中，依照 3.3 提到的方法透過 588 筆商品 D'' 從 836 筆商品 D''' 提取了 1,613 筆三元組樣本。再來將其拆解為 1,613 筆 (Anchor, Positive, "match") 和 1,613 筆 (Anchor, Negative, "no-match") 的數據，並將其透過 3.3 的方法抽取 3,226 筆負樣本，最終獲得了 6,452 筆 Match 模型的測試資料，組成了 B2C_Match_Test 資料集。

B2C_ Test：最後是我們的兩階段模型評估的測試集 B2C_Test。我們透過兩個商品集 D'' 和 D''' 來製作我們的兩階段評估測試集 B2C_Test。首先我們從商品集 D'' 中取出一個商品 $e_i^3 \in D''$，然後標註從 $e^4 \in D'''$ 找出所有相同的商品，組成(Anchor,Positive,"match")，而剩下不同的商品則組成 (Anchor, Positive,"no-match")，將此流程經過 i 次就能得到所有的測試資料。

由於我們的 D'' 總共有 588 筆資料，而 D''' 有 836 筆資料，於是就能夠得到 588*836=491,568 筆測試兩階段績效的資料，也就是 B2C_Test 資料集。資料型格式為 $\{(e^3, e^4, l) | l \in (\text{"match"}, \text{"no-match"})\}$。整體的資料集如表 1 所示 (見第 2 頁)。

4.3 實驗結果

本章節旨在呈現不同模型在商品比對任務中的表現。我們針對 B2C 的情境並評測第一階段中不同的 K 值對於模型的影響，以及使用兩階段和單一階段模型的運算時間和準確度，藉此驗證兩階段的必要性。

資料集名稱	資料數量	資料格式
T("PChome", "Momo")	Root :3,526 Leaf:4,165	$(D\ U\ D')$, $(D''\ \mathrm{U}\ D''')$, $T\ =\ \{(e^1, e^2, \text{"}match\text{"})\}$
B2C_block_Train	4,088 筆	D, D' Triplet (Negative, Anchor, Positive)
B2C_block _Test	588 個 Query 836 個 Leaf	D'', D''' 關聯資料 (Product Name, Connect)
B2C_match _Train	16,352 筆	D, D' 分類標籤 (Anchor, Positive, "match") U (Anchor, Negative, "no-match")
B2C_match_Test	6,452 筆	D'', D''' 分類標籤 (Anchor, Positive, "match")
B2C _Test	491,568 筆	D'', D''' $\{(e^3, e^4, l) \mid l \in (\text{"}match\text{"}, \text{"}No\_match\text{"})\}$

表 1. 針對不同任務的資料集

B2C 兩階段不同預訓練模型實驗：在這個實驗中，我們比較由中文新聞和維基百科資料訓練的 CKIP-BERT 和我們自己透過商品名稱訓練的 eComBERT 作為預訓練模型於兩階段模型的效果。

我們使用 B2C_Block_Train 資料集作為訓練第一階段 Blocking 模型的資料集，並使用 C2C_Match_Train 資料集作為第二階段 Matching模型的訓練資料，最後我們將其測試在 B2C_Test 的資料集上，在這裡第一階段過濾的商品數量 $K = 50$，實驗結果如下表 2 所示。

	Accuracy	Recall	Precision	F1
eCom BERT	0.942	0.981	0.872	0.923
CKIP BERT	0.928	0.827	0.735	0.778

表 2. B2C 兩階段不同預訓練模型實驗數據表

B2C Blocking 不同預訓練模型實驗：我們針對第一階段的 Blocking 模型使用 B2C_Block_Train 訓練，在這個實驗中我們主要關注 Recall@K 的指標，因為在此我們更注重模型是否能幫我們找出實際上相同的商品。同樣的我們也會比較不同預訓練模型的效果。實驗結果如下表 3 所示。

Recall	@1	@5	@10	@20	@50
eComBERT	0.765	0.851	0.893	0.927	0.998
CKIP BERT	0.635	0.824	0.879	0.939	0.996

表 3. B2C Blocking 不同預訓練模型實驗數據表

B2C Matching 不同預訓練模型實驗：在第二階段的 B2C Matching 模型實驗中，我們使用 B2C_Match_Train 進行訓練，並用 B2C_Match_Test 測試模型，在這裡我們也比較使用不同預訓練模型的效果。實驗結果如下表 4 所示。

	Accuracy	Recall	Precision	F1
eCom BERT	0.982	0.992	0.876	0.930
CKIP BERT	0.926	0.974	0.758	0.853

表 4. B2C Matching 不同預訓練模型實驗數據表

兩階段不同 K 值對比實驗：

在第一階段中過濾的商品數量 K 會對第二階段的效能產生顯著影響，因此我們針對不同的 K 值進行實驗，這些 K 值代表第一階段過濾後傳遞給第二階段的商品數量，我們將我們訓練的模型測試在 B2C_Test 測試集上，其結果如下表 5 所示。

	Accuracy	Precision	Recall	F1
K=10	0.946	0.976	0.873	0.922
K=50	0.942	0.981	0.872	0.923
K=100	0.945	0.943	0.853	0.896

表 5. 兩階段不同 K 值對比實驗數據表

兩階段與單階段績效對比實驗：商品匹配的方法可分成以下三種，分別是：(1) Only Similarity：將所有的商品透過 BERT 模型轉換成向量之後，透過相似度計算來尋找相似商品，並使用閥值來判定是否為相同商品，該方法只需要將所有商品推論一次即可；(2) Only Classify：將兩個商品名稱分別作為

BERT 的輸入，並利用輸出 [CLS] Token 通過分類器來判斷商品是否相同，此方法需要對所有商品進行 BERT 模型的推理；(3) Two Stage：本論文採用的兩階段匹配方法。

由於本論文所採用兩階段方法的原因主要是計算時間以及匹配精確度的協調。為此，我們分別針對使用相似度、分類器以及兩階段的作法，計算單一個商品 $e_i^1 \in D$ 在 10,000 筆的 $e^2 \in D'$ 中尋找並辨識相同商品所需的時間，以及利用 B2C_Test 測試集測試所得的 F1 Score 分數。下表 6 為對比數據表。

	Time Type	10,000 Products	F1 Score
Only Similarity	Inference Time	35.091s	86.709
	Similarity Dot Product Time	2.061s	
Only Classify	Inference Time	1002.971s	95.231
Two Stage	Inference Time	19.353s	93.893
	Similarity Dot Product Time	0.2087s	
	Classify Time	11.7073s	

表 6.兩階段與單階段績效對比實驗對比表

實驗結果分析：根據實驗結果，我們可以明確地發現，基於 eComBERT 微調的模型在效能上超過基於 CKIP BERT 微調的模型許多。因此對於不同領域的 In-Domain 預訓練對於模型的下游任務尤為重要。此外，針對兩階段的方法，雖然單純使用分類模型的效果最佳，但運算時間非常長，無法滿足商品比對的即時性需求。而僅使用語意相似度的搜尋方法，雖然速度最快，但效果明顯不如兩階段方法。綜合來看，我們選擇的兩階段方法能夠在相對較短的時間內，提供效果適中的解決方案，是較為理想的選擇。

5 結論

在電子商務領域，構建精準的產品匹配模型需要大量高質量的訓練數據，但現有資料集無法滿足台灣電商市場的需求。為解決這一問題，我們設計了一個專門用於電商商品收集與標註的平台，可以自動化地從不同電商平台搜尋並對比商品，進行精確的匹配與標註，生成有效的正負樣本。

雖然目前已有多種基於不同語言和領域的 BERT 預訓練模型，但它們不完全適用於台灣電商環境，例如 CKIP-BERT 使用通用中文語料進行預訓練，無法捕捉電商文本中的專業特徵。因此，我們使用 2.6 億筆中文電商文本進行重新預訓練，提出了針對電商領域優化的預訓練模型 eComBERT。

基於 eComBERT，我們進一步設計了一個專為電商商品匹配問題量身打造的兩階段架構 eComMatch。該架構包含兩個模組：在第一階段，我們採用結合 Triplet 結構的 BERT 模型進行初步過濾，快速篩選潛在匹配商品；在第二階段，我們利用 Siamese 網絡結構進行精確匹配，判斷篩選後商品之間的相似性。這樣的設計提升了匹配效率與準確度，為電商商品匹配領域的研究與應用奠定了堅實的基礎。

References

Chopra, S., Hadsell, R., & LeCun, Y. (2005). *Learning a similarity metric discriminatively, with application to face verification.* Paper presented at the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05).

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hoffer, E., & Ailon, N. (2015). *Deep metric learning using triplet network.* Paper presented at the Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3.

Lin, Z., & Chen, W. (2022). *ALFA Matcher: Supervised and Unsupervised Product Entity Matching Model.* Paper presented at the 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA).

Mihalcea, R., & Csomai, A. (2007). *Wikify! Linking documents to encyclopedic knowledge.* Paper presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., . . . Raghavendra, V. (2018). *Deep learning for entity matching: A design space exploration.* Paper presented at the Proceedings of the 2018 international conference on management of data.

Peeters, R., & Bizer, C. (2022). *Supervised contrastive learning for product matching.* Paper presented at the Companion Proceedings of the Web Conference 2022.

Rao, D., McNamee, P., & Dredze, M. 2013. Entity linking: Finding extracted entities in a knowledge base. *Multi-source, multilingual information extraction and summarization*, 93-115.

Reimers, N. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.

Shah, K., Kopru, S., & Ruvini, J. D. (2018). *Neural network based extreme classification and similarity models for product matching.* Paper presented at the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers).

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, J., Li, Y., & Hirota, W. (2021). *Machamp: A generalized entity matching benchmark.* Paper presented at the Proceedings of the 30th ACM International Conference on Information & Knowledge Management.

Yuan, H., Liu, G., Li, H., & Wang, L. (2018). *Matching recommendations based on siamese network and metric learning.* Paper presented at the 2018 15th International Conference on Service Systems and Service Management (ICSSSM).

基於社群媒體情緒與圖神經網路進行股票趨勢預測
Stock Trend Prediction with Social Media Sentiment and Graph Neural Network

Yen-Tsang Wu　Jenq-Haur Wang
Department of Computer Science and Information Engineering
National Taipei University of Technology
Web Information Retrieval Lab
buddyswu@gmail.com
jhwang@ntut.edu.tw

Po Chuan Huang
Taiwan Semiconductor Manufacturing Company Limited
peter02589@gmail.com

摘要

本文提出一種結合歷史股價與社群情緒的股票漲跌預測多模態整合架構。首先，我們將過去幾天的移動平均使用 GRU 取得序列式特徵，再將過去幾天股價之相對價差透過圖注意力機制得出價差特徵，並將社群言論之情緒分析也透過圖注意力機制得出情緒特徵；我們將三種不同性質之特徵互相結合，透過圖注意力機制得出股票特徵。最後將不同股票透過超圖神經網路預測出股票漲跌之結果。實驗結果顯示，本論文提出的模型在結合了多種不同性質特徵與考量不同股票之間關係後，相較於先前的方法，能更有效的偵測出股票漲跌。

Abstract

This paper presents a novel multimodal fusion approach for stock trend prediction, integrating historical stock prices and social sentiment data. The proposed method first extracts sequential features from stock price moving averages over recent days using a Gated Recurrent Unit (GRU). Simultaneously, it captures relative spread features by applying Graph Attention Networks (GAT) to the relative price spreads, and sentiment features are derived from social sentiment analysis, also using GAT. These three feature sets are then fused using a graph attention mechanism to obtain comprehensive stock representations. Subsequently, a stock correlation graph is constructed, where nodes represent individual stocks, and edges reflect their correlations. Hypergraph neural networks are employed to predict stock trends based on this graph structure. Experimental results demonstrate that the proposed method achieves an accuracy of 0.604. By incorporating diverse feature types and accounting for inter-stock relationships, this approach significantly outperforms previous models in predicting stock price movements.

關鍵字：圖神經網路、社群媒體情緒、股票趨勢預測

Keywords: Graph neural network, Social media sentiment, Stock trend prediction

1　介紹

近年來，隨著經濟發展和全球化，股票市場交易活躍，股票投資成了重要的經濟行為。由於股票市場是一個複雜的系統，受到多種因素的影響，傳統的股票預測方法主要依賴於歷史股價數據的分析，如價格走勢圖、交易量等些指標能夠反映出市場的整體走向和投資者的交易行為 (Duan et al., 2022; Zhang et al., 2017)。
隨著社群媒體影響力日益增強，投資者情緒和觀點越來越多地在網路上表達。眾多研究表明，社交媒體上的情感信息可以反映市場情緒的變化，並影響股票價格的走向 (Behrendt et al., 2018; Ben Cheikh et al., 2024)。學者們指出，僅依靠分析社交媒體上與股票相關的討論和情感提供的額外信息，就能有助於預測未來的價格走勢。Tabari 等人 (Tabari et al., 2018) 研究發現,平均而言,社交媒體上發文的情緒會導致當天股票收益上升 0.26%。除此之外，由於美國股市在全球交易中占據重要地位， Mora 等人 (Nuñez-Mora et al., 2023) 分析了約 5000 萬條推文，計算了 2557 家美國上市公司的正面和負面情緒因子，結果顯示，503 家公司的負面情緒因子影響大於正面情緒，證明社交媒體情緒已成為市場的一個影響源。多項研究都支持社交媒體情緒能夠反映市場

情緒變化，並影響股票價格走向的觀點。投資者可利用社交媒體情緒信息來預測股價。Wang 等人 (Wang et al., 2023) 提出一個名為 ECON 的框架，利用推文、宏觀經濟指標和歷史價格來預測股票運動和波動性。該研究強調了推文數據的質量和行業相關性對預測的影響。Asgarov 等人 (Asgarov, 2023) 評估了社交媒體情感在預測主要公司股價中的有效性，他們使用 LSTM 模型分析推文情感和歷史價格數據，結果顯示情感表達與股價波動之間存在強相關性。Amin 等人 (Amin et al., 2024) 探討了社交媒體上關於人工智慧進展的情感是否能預測相關公司的日常股價波動。研究利用自然語言處理技術分析推文情感，並使用多種機器學習模型進行預測，發現推文情感與市場價值之間存在潛在的相關性。

基於上述的研究，我們提出一個結合技術指標、相對價差與社群言論情緒，並使用圖神經網路融合多種特徵，最後使用 Hypergraph Neural Networks (Feng et al., 2019)結合股票之間的關聯性，結合一個多層次圖神經網路模型，以提升股票趨勢預測的準確性。我們的貢獻如下：1.結合歷史的股價、股價之間的變化、社群平台情緒三種資訊來進行股票趨勢預測 2.提出一個多項特徵結合圖神經網路的融合技術。

2 相關研究

本章節介紹股票趨勢預測的相關研究，主要分為基於股價、社群平台、以及結合兩者資訊的研究，並探討不同特徵下使用的遞歸神經網路和圖神經網路等模型架構技術。

2.1 基於序列式神經網路模型股票趨勢預測

2017 年由 Selvin 等人 (Selvin et al., 2017) 使用每日收盤價預測單間公司的股票走勢，他們使用三種不同的深度學習模型 RNN、CNN、LSTM 來進行訓練與預測。2020 年 Lu 等人 (Lu et al., 2021) 提出 CNN-BiLSTM-AM 方法，以歷史價格、成交量、漲跌變化作為輸入，由卷積神經網路(CNN)、雙向長短期記憶神經網路(BiLSTM)和注意力機制(AM)組合，CNN 用於提取輸入數據的特徵，BiLSTM 用於學習和預測提取的特徵數據，而 AM 用於捕捉過去不同時間的特徵對第二天收盤價的影響。2024 年 LI 等人 (Li et al., 2024) 提出 MASTER 模型利用市場訊息進行自動特徵選擇，以適應市場的動態變化。MASTER 能夠有效處理複雜的時間序列資料，提高股價預測的準確性。

2.2 基於社群平台言論股票趨勢預測

Nguyen 等人 (Nguyen & Shirai, 2015)於 2015 年提出 TSLDA 模型，透過社群平台上的言論進行情緒分析，並將其輸入至主題模型以捕捉相關主題，用於預測股價走勢。2018 年由 Hu 等人 (Hu et al., 2018)提出一個多層注意力網路模型，將單日的發文輸入至注意力網路以獲得其表示形式，然後將多日的發文表示形式輸入至序列式神經網路模型與注意力網路，用於預測股票趨勢。同年 Xu 等人 (Xu & Cohen, 2018) 提出 StockNet 模型，同時結合了股價特徵與 tweets 文本特徵，其中股價特徵使用序列式神經網路模型，而文本特徵亦是使用多層注意力網路進行編碼，並且比較了五種模型組合的表現。2024 年 Fan 等人 (Fan & Shen, 2024) 提出一種基於多層感知機（MLP）的簡單而強大的股價預測架構。該模型通過三種混合機制有效地捕捉股票數據中的複雜相關性。同年，Ranjith 等人 (Ranjith, 2024) 提出了一種新穎的可解釋人工智慧（XAI）模型，該模型結合了多種數據源，包括社交媒體情緒和技術指標，以預測股票市場趨勢並提供解釋性結果。

2.3 基於圖神經網路進行股票趨勢預測

近年許多研究提出使用圖神經網路來學習股票間複雜的關聯性，Chen 等人 (Chen et al., 2018) 在 2018 年提出了一種基於 LSTM+GCN 的模型，該模型使用歷史股價作為節點表示形式，並通過圖卷積神經網路（GCN）來學習股票之間的關係，進行訓練以預測股票趨勢。2019 年 Kim 等人 (Kim et al., 2019b)提出 HATS 模型，該模型將每檔股票的歷史股價作為輸入，並通過圖注意力機制聚合不同關係類型的股票表示形式。HATS 不僅用應用於預測個股價格，還可用於預測市場指數的走勢。2020 年由 Wei Li 等人 (Li et al., 2021)提出一種基於 LSTM-RGCN 模型用於股票趨勢預測。由於股票在休市期間沒有交易操作，休市後的新聞往往會對下個開盤日的股市產生影響。因此，該研究根據股市收盤期間的新聞來預測下個開盤日股票走勢。該模型結合了財經

新聞文本與歷史股價當作為特徵，使用 LSTM 學習，並通過 GCN 捕捉學習股票之間的依賴性，最終預測下個開盤日股票走勢。2020 年由 Sawhney 等人 (Sawhney et al., 2020)提出一種名為 MAN-SF 的模型架構，從歷史股價、社群平台、股票間關係中進行聯合學習，具體而言，模型通過 GRU 捕捉歷史股價的數值型時間序列關係與社群平台的文本上下文關係，並進一步通過圖神經網路學習股票之間的相互關係。該模型在真實世界的股市數據上進行了實驗，驗證了其有效性。

由於股票預測具有高度複雜性，部分研究選擇使用超圖神經網路模型來學習，2021 年 Sawhney (Sawhney et al., 2021)等人提出 STHAN-SR，一種用於選擇股票的模型。該模型通過LSTM與注意力機制的組合來提取股價的時間序列特徵，並使用圖模型學習股票之間的複雜依賴關係。最終，根據每檔股票的預期利潤對其進行排名。

3 方法

本研究提出的架構分為三個部分，包括特徵提取（Feature Extraction）、模型訓練（Model Training）和分類（Classification）。特徵提取負責擷取數據特徵與情緒特徵；模型訓練部分結合序列式神經網路模型與圖神經網路模型，針對不同性質的特徵進行相應的模型訓練；最後，分類部分將模型輸出的向量進行分類處理。架構圖如圖 1 所示。

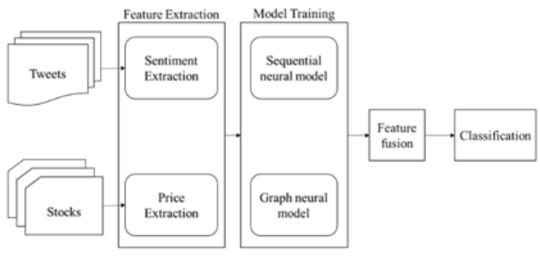


圖 1.系統架構圖

3.1 特徵擷取

特徵擷取分成股價特徵跟情緒特徵兩個部分。

3.1.1 股價特徵擷取

股價特徵由每檔股票每日的日期、開盤價、收盤價、最高價、最低價及成交量構成。我們使用還原K線的收盤價來計算股價的漲跌情況。定義如下：若當天 t 的收盤價大於或等於前一天 t-1 的收盤價，則當天 t 被定義為「漲」；若當天 t 的收盤價小於前一天 t-1 的收盤價，則當天 t 被定義為「跌」，如公式(1)所示。公式中的 p_t^{ac} 代表某檔股票在第 t 天的股價，其中 ac 表示還原 K 線的收盤價。此時，若 $y_t = 0$，則表示第 t 天股票下跌；若 $y_t = 1$，則表示第 t 天股票上漲。

$$y_t = \begin{cases} 0, p_t^{ac} < p_{t-1}^{ac} \\ 1, p_t^{ac} \geq p_{t-1}^{ac} \end{cases} \quad (1)$$

根據 Sawhney 等人的研究，決定股票趨勢的關鍵在於價格變化，而非價格本身。雖然每日的股價確實具有重要意義，但真正引發股勢變化的是歷史股價的價差。因此，我們採用由開盤價、收盤價、最高價和最低價計算出的四種價差值作為特徵，具體如公式 (2) 所示。

$$c_t = p_t - p_{t-1} \quad (2)$$

公式中的 p_t 代表某檔股票在第 t 天的股價，將其與前一天 t-1 的股價相減，便可得出價差 c_t，相同的作法，使用開盤價得到開盤價價差 c_t^o、使用最高價得到最高價價差 c_t^h、使用收盤價得到收盤價價差 c_t^c、使用最低價得到最低價價差 c_t^l。此外，觀察股勢變化的另一重要指標是移動平均線，其目的是通過計算過去價格的平均值來平滑化價格波動，使價格走勢更為明確。因此，我們使用收盤價來計算五天、十天、二十天及六十天的移動平均，如公式 (3) 所示。

$$m_t^T = \frac{\sum_{i=t-T+1}^{t} p_i^c}{T} \quad (3)$$

公式中的 p_i^c 代表某檔股票在第 i 天的收盤價，T 為觀察的過去天數。通過將過去 T 天的收盤價總和除以 T，即可得到第 t 天的 T 天移動平均值 m_t^T。而五天的移動平均值表示為 m_t^5，十天的移動平均值為 m_t^{10}，二十天的移動平均值為 m_t^{20}，而六十天的移動平均值則為 m_t^{60}。

3.1.2 情緒特徵擷取

首先，我們對原始推文移除 URL 及停用字，然後在文本中出現的股票代碼與股票名稱處加上[ASP]目標標籤，接著，我們使用 Feng 等人 (Feng et al., 2019) 提出的 ABSC 模型進行情緒辨識。該模型並不是將整篇推文進行整體情緒評估，而是針對推文中每一個目標標籤進行情緒分類，情緒類別包括負面（Negative）、中性（Neutral）和正面（Positive）。我們利用 Word2vec (Mikolov,

2013) 將這三種情緒類別轉換為嵌入向量（embedding），作為情緒特徵向量。資料集中所使用的股票均屬於高討論度金融商品，因此，我們將單一股票在一天內的推文情緒特徵向量相加並取平均，作為該天的情緒特徵向量。此外，為了讓情緒較為明顯的向量擁有較大的影響力，我們對負面情緒和正面情緒向量賦予不同的權重，如公式 (4) 所示。

$$e_t = e^{Positive} \times \frac{n_t^{Positive}}{n_t^{Positive} + n_t^{Negative}} + e^{Negative} \times \frac{n_t^{Negative}}{n_t^{Positive} + n_t^{Negative}} \quad (4)$$

$e^{Positive}$ 代表正面情緒向量，$e^{Negative}$ 代表負面情緒向量，$n_t^{Positive}$ 為某檔股票在第 t 天的正面情緒發文數。$n_t^{Negative}$ 為某檔股票在第 t 天的負面情緒發文數。通過此公式進行加權相加，我們可以得出該檔股票在第 t 天的情緒特徵向量 e_t。

3.2 架構

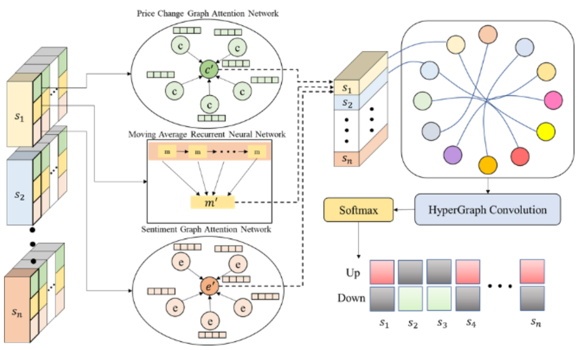


圖 2. CME-HG 模型架構圖

如圖 2 所示，我們提出了一種名為 CME-HG 的預測模型。並定義每檔股票 $s \in S = \{s_1, s_2, \ldots, s_n\}$，其中，對於每檔股票 s_i，所有天數的特徵按以每 T 天劃分為一個資料樣本。每個資料樣本包含三種類型的特徵：

1. 價差特徵 $C_t = \{c_{t-T+1}, c_{t-T+2}, \ldots, c_t\}$：特徵由每日的價格變動計算得出，旨在捕捉價格波動情況，進而預測未來趨勢。
2. 移動平均特徵 $M_t = \{m_{t-T+1}, m_{t-T+2}, \ldots, m_t\}$ 此特徵通過移動平均來平滑價格變化，以觀察短期和長期趨勢變化。
3. 情緒特徵 $E_t = \{e_{t-T+1}, e_{t-T+2}, \ldots, e_t\}$：此特徵來自於社交媒體情緒的變動。藉以捕捉情緒變化對股價的影響。

這三種特徵通過各自專屬的模型訓練方式進行處理，然後再進行融合，以提升預測的準確性。

3.2.1 個股的價格變化特徵

我們定義了一個「Price Change-to-Price Change」同構圖 $G = (V, E)$，其中 V 和 E 分別代表圖的節點與邊。節點 V 由第 t 天的價差特徵 c_t 組成。而價差特徵 c_t 包含四個部分：$[c_t^o, c_t^h, c_t^c, c_t^l]$，分別代表開盤價、最高價、收盤價和最低價的價差，這些組成作為節點的初始向量。為了得到過去一段時間的平均價差值，我們將過去 T 天內的價差特徵中的各項分別相加並取平均，得到一個平均價差特徵向量 c_{t+1} 作為新的節點初始向量。這樣的平均處理能夠平滑化數值波動，使數據更具穩定性，如公式 (5) 所示。

$$c_{t+1} = \frac{\sum_{i=t-T+1}^{t}[c_i^o, c_i^h, c_i^c, c_i^l]}{T} \quad (5)$$

每一天的價差特徵向量包括四個值，分別是根據開盤價計算得到的 c^o、最高價計算得到的 c^h、收盤價計算得到的 c^c 以及最低價計算得到的 c^l。我們將過去 T 天中這四個價差值分別相加並取平均，從而得出第 t+1 天的平均價差特徵向量 c_{t+1}。

邊 E 由 $E_{c_i c_j}$ 表示，具體來說，$E_{c_t c_{t+1}}$ 表示第 t 天的價差特徵 c_t 與第 t+1 天的平均價差特徵向量 c_{t+1} 之間的關係。這用來表示過去的價格波動可能會影響到未來投資者的決策。因此，我們將過去 T 天的價差特徵與第 t+1 天的價差特徵逐一建立邊，如圖 3 所示。

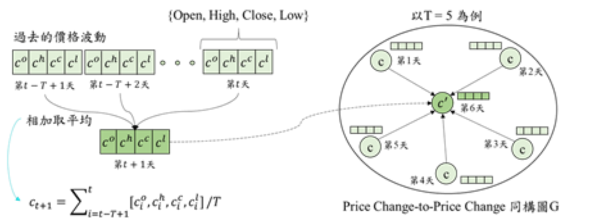


圖 3. Price Change-to-Price Change 同構圖 G 說明

我們將有向圖 G 作為輸入，並使用圖注意力機制網路模型（Graph Attention Network, GAT），通過引入注意力機制（attention mechanism）。對於圖中的一對節點 (i, j)，，我們使用自注意力機制來學習注意力係數 e_{ij}，該係數表示節點 i 對節點 j 的重要性，如公式 (6) 所示

$$e_{ij} = attention(Wc_i, Wc_j), c_i, c_j \in C_t \quad (6)$$

我們將 c_i 與 c_j 進行線性轉換，然後將兩者串接起來，接著乘以上轉置的權重向量 a^T。隨後，結果會輸入至 LeakyReLU 激活函數，並通過 softmax 函數進行正規化，從而得到注意

力權重。N_i代表節點 i 的鄰居節點，這裡指的是該節點前五天的節點（即 i-5, ..., i-1）。當計算出每一對節點之間的權重 $\alpha_{i,j}$ 後，我們將所有鄰居節點的特徵乘上相應的權重，再加總得到更新後的節點特徵表示 c'_{t+1}，如公式 (7) 所示。

$$c'_{t+1} = \sigma(\sum_{j0\in\mathcal{N}_i} \alpha_{i,j} W c_j) \quad (7)$$

最後，我們採用多頭注意力機制來學習更穩定的嵌入表示（embedding representation）。如公式 (8) 所示，我們將公式 (7) 的轉換過程執行 k 次，並將每次輸出的表示 c'_{t+1} 串接起來，從而獲得最終的輸出表示 c'_{t+1}。這樣的多頭注意力機制能夠捕捉更多樣的關係和特徵，從而提高模型的表現穩定性和泛化能力。

$$c'_{t+1} = \overset{k}{\underset{k=1}{\parallel}} \sigma(\sum_{j\in\mathcal{N}_i} \alpha_{i,j}^{k} W^{k} c_j) \quad (8)$$

3.2.2 個股移動平均序列式特徵

我們將過去 T 天的移動平均特徵作為輸入，並使用門控循環單元（GRU）來提取移動平均線的時間序列特徵。第 t 天的 GRU 輸出表示如公式 (9) 所示。這樣可以有效捕捉移動平均線在時間序列中的動態變化，進而提升對股價趨勢的預測能力。

$$h_t = \mathrm{GRU}(m_t, h_{t-1}) \quad (9)$$

我們定義第 t 天的移動平均特徵為m_t，由$m_t^5, m_t^{10}, m_t^{20}, m_t^{60}$組成，其中 m_t^5代表五日線，m_t^{10}代表十日線，m_t^{20}代表二十日線，m_t^{60}代表六十日線。為了更有效地捕捉不同天數移動平均線對股價趨勢的影響，我們使用時間注意力機制來學習不同天數之間的移動平均特徵對預測結果的影響權重，並對 GRU 的所有隱藏層特徵進行聚合。權重的計算方式如公式 (10) 所示。

$$\alpha_i = \frac{\exp(h_i^T W \bar{h}_z)}{\sum_{i=1}^{T} \exp(h_i^T W \bar{h}_z)} \quad (10)$$

其中，$\bar{h}_z$代表 GRU 的過去隱藏層表示，α_i 為第 i 天的注意力權重，W 為一個可學習的參數矩陣。移動平均線將被輸入至 GRU 和時間注意力機制，具體結構如圖 4 所示。這樣的設計能夠充分利用時間序列的特徵，進一步提高模型對股價趨勢預測的準確性。

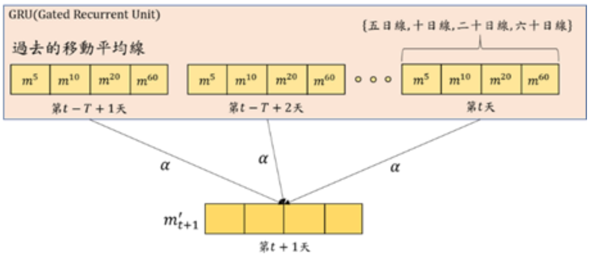


圖 4. 移動平均線輸入 GRU 訓練示意圖

3.2.3 個股評論情緒特徵學習

情緒特徵的處理方法與價差特徵的訓練方式相似。我們定義 sentiment-to-sentiment 同構圖 G=(V,E)，其中 V 和 E 分別代表圖的節點和邊。節點 V 由第 t 天的情緒特徵 e_t 組成，而情緒特徵 e_t 由使用 word2vec 模型轉換而成的向量組成，這些向量作為節點的初始表示。為了獲得過去一段時間的平均情緒特徵向量，我們將過去 T 天的情緒特徵向量e_i相加並取平均，從而得到第 t+1 天的平均情緒特徵向量 e_{t+1}，該向量將作為節點的初始向量，如公式 (11) 所示。

$$e_{t+1} = \frac{\sum_{i=t-T+1}^{t} e_i}{T} \quad (11)$$

邊 E 由 $E_{e_i e_j}$表示，其中$E_{e_t e_{t+1}}$描述第 t 天的情緒特徵e_{t+1}與第 t+1 天的平均情緒特徵向量 e_{t+1} 之間的關係，如圖 5 所示。

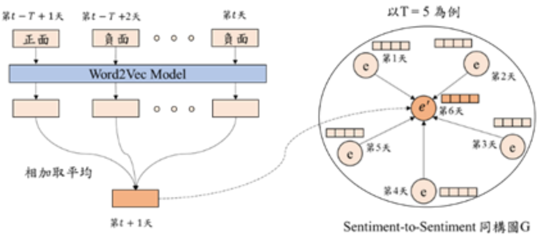


圖 5. Sentiment-to-Sentiment 同構圖 G 說明

3.2.4 特徵融合圖神經網路

定義特徵融合圖 G=(V,E)，其中 V 和 E 分別作為圖的節點與邊。節點 V 由第 t+1 天的價差特徵c'_{t+1}、移動平均特徵 m'_{t+1}、情緒特徵 e'_{t+1}組成，此外還包含一個融合節點 f_{t+1}。該融合節點f_{t+1}是通過將上述三個特徵向量相加得到的。邊 E 則連接這三個特徵向量節點與融合節點f_{t+1}，用以表示不同性質的特徵與融合節點之間的關係，如圖 6 所示。

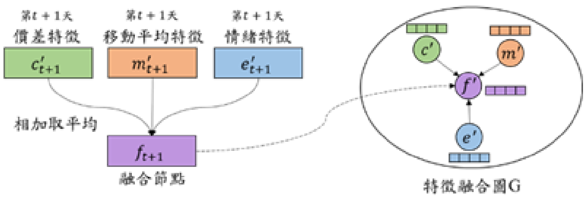


圖 6. 特徵融合圖說明

接著，我們將特徵融合圖 G 輸入至圖注意力網路（Graph Attention Network, GAT）進行訓練。訓練完成後，從網路中提取融合向量 f'_{t+1}，該向量作為每檔股票在不同性質特徵融合後的股票表示形式。

3.2.5 不同股票間關係特徵

我們建構 stock-to-stock 的同構圖，其中節點為每檔股票的表示形式，邊則依據 Feng 等人 (Feng et al., 2019) 提出兩種關係來構建不同股票之間的關聯，第一種關係是**行業關係**，根據全球行業分類標準 (Global Industry Classification Standard, GICS)，我們將屬於同一行業的股票進行連邊。在我們研究的 85 檔股票中，共構建了 16 種行業關聯。例如，Computer Software 這一行業包含 Google 和 Facebook，這兩家公司作為同一行業的股票因此相連。第二種關係是基於 **Wikidata** 提供的股票資訊。例如，Alphabet Inc. 是 Google LLC 的母公司，而 Microsoft 與 Branded Entertainment Network 由共同的董事 Bill Gates 連接。在 85 檔股票中，共構建了 70 種這樣的關聯。通過這兩種關係，我們最終構建了 86 種股票之間的關聯，具體示意圖如圖 7 所示。

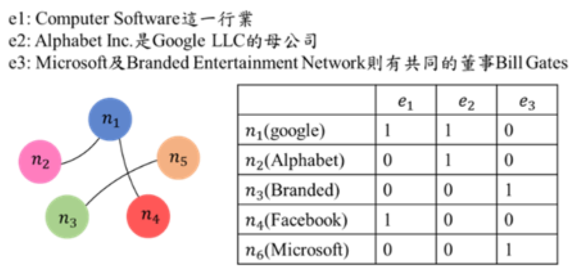


圖 7. HyperGraph 建圖示意圖

由於我們所建構的關聯來自不同概念，包括同一行業的關聯、母公司與子公司的關聯、以及同一董事的關聯，這些關聯的複雜性較高。基於此，我們採用 Feng 等人 (Feng et al., 2019) 提出的 HyperGraph Neural Networks 來處理這種更為複雜的高階關聯性，並學習股票的表示形式。如公式 (12) 所示，H 代表 85 檔股票所構建的 86 種關聯的矩陣，$f^{(l)}$ 為第 l 層的股票表示形式：

$$f^{(l+1)} = \sigma\left(D_v^{-\frac{1}{2}} H W D_e^{-1} H^T D_v^{-\frac{1}{2}} f^{(l)} \Theta^{(l)}\right) \quad (12)$$

經過超圖神經網路的訓練，我們獲得每檔股票的二維向量表示，並通過 softmax 函數對結果進行正規化，從而得到預測結果 p_t^k，表示第 k 檔股票在第 t 天的預測結果。為了計算分類結果的誤差，我們使用交叉熵損失函數 (cross entropy)，如公式 (13) 所示，其中 n 為股票的數量，y_t^k 為第 k 檔股票在第 t 天的實際漲跌結果：

$$\mathcal{L} = \sum_{k=1}^{n} -[y_t^k * \log(p_t^k) + (1 - y_t^k) * \log(1 - p_t^k)] \quad (13)$$

最後，我們對所有股票在所有測試天數內的預測結果進行加總並取平均，從而得到最終的預測結果。

4 實驗

我們將在以下章節詳細描述實驗參數與結果。

4.1 資料集

本文使用了由 Yum 等人 (Xu & Cohen, 2018) 在 2018 年提供的結合股票與推文進行股票預測的數據集。該數據集包含 88 檔具有高討論度的股票，均來自美國的 S&P 500 和 NASDAQ 指數，數據範圍為 2014 年 1 月 1 日至 2016 年 1 月 1 日。扣除非交易日後，總共有 504 天的交易記錄，88 檔股票共計 41,990 筆數據，其中股價上漲的筆數為 21,679，股價下跌的筆數為 20,311，漲跌比例分別為 52% 和 48%。此外，該數據集中還包含了 2014 年 1 月 1 日至 2016 年 1 月 1 日期間提及這些股票的推文，共計收集了 106,338 篇推文。由於 BABA (Alibaba Group Holding)、AGFS (AgroFresh Solutions Inc.) 和 GMRE (Global Medical REIT) 這三檔股票沒有提供相關推文，因此我們將其從數據集中移除。數據集的劃分比例為 7:1:2，具體的日期範圍如表 1 所示。

	Training	Validation	Testing
日期	2014/1/1-2015/7/31	2015/8/1-2015/9/30	2015/10/1-2016/1/

表 1.資料集日期

4.2 實驗參數

實驗環境基於 Ubuntu 18.04 操作系統，運行於具有 32GB 記憶體的主機上，顯卡為 GeForce RTX 2080 Ti（11GB）。實驗使用 Python 3.8.3 和 PyTorch 1.10.2+cu102 進行。實驗中，價差特徵與情緒特徵的計算均使用 GAT，隱藏層向量的維度為 4，Attention head 設為 1。Word2vec 的 embedding 維度設為 300。移動平均特徵則是通過 GRU 模型計算，隱藏層向量的維度同樣為 4。CME-HG 模型使用的損失函數為 Cross Entropy Loss，dropout 設定為 0.38，Optimizer 為 Adam(Diederik, 2014)，Learning Rate 設為5×10^{-5}，最大訓練 epoch 數為 200。

4.3 評估指標

實驗評估指標採用兩種評估方式，分別為 Accuracy 與 F1-Score。這兩項指標能夠有效衡量模型在分類任務中的表現，其中 Accuracy 用於評估模型的整體正確率，F1-Score 則綜合考慮了 Precision 與 Recall。

4.4 比較模型

為了驗證所提出模型的效能，我們參考 MAN-SF (Sawhney et al., 2021) 的方法，使用以下模型作為比較對象：

- ARIMA (Autoregressive Integrated Moving Average model) (Brown, 2004)：：也稱為整合移動平均自我迴歸模型，這是一種時間序列預測方法，使用歷史股價進行預測。
- Selvin et al. (Yang et al., 2021)：分別使用三種深度學習模型 RNN、CNN 和 LSTM。我們選擇表現最佳的 LSTM 作為比較模型。
- RandForest (Breiman, 2001)：將推文文本經由 word2vec 轉換後，使用 Random Forests classifier 進行訓練與分類。
- TSLDA (Nguyen & Shirai, 2015)：使用情緒分析和主題模型來分析社群媒體言論，並進行股票趨勢預測。
- HAN (Hu et al., 2018)：利用分層注意力機制，對單日與多日的 tweets 文本進行編碼，以預測股票趨勢。
- StockNet (Xu & Cohen, 2018)：使用股價與 tweet 文本特徵，通過分層注意力對單日與多日的 tweet 進行編碼，並使用序列式神經網路模型處理價格特徵。
- Chen et al. (Chen et al., 2018)：該模型使用歷史股價作為節點表示形式，並通過 GCN 整合公司之間的關係進行股票趨勢預測。
- HATS (Kim et al., 2019a)：該模型使用多檔股票的歷史價格特徵作為節點表示形式，並通過圖注意力機制結合每檔股票之間的關係進行股票趨勢預測。

4.5 實驗結果

實驗結果如表 2 所示。表格中展示了不同模型的比較結果，包括準確率（Accuracy）、F1-Score 等性能指標。

Method	Model	Accuracy	F1-score
Regression	ARIMA	0.514	0.513
RNN	Selvin et al	0.530	0.529
Social Media	RandFores	0.531	0.527
	TSLDA	0.541	0.539
	HAN	0.576	0.572
RNN+ Attention	StockNet	0.550	0.546
Graph	Chen et al	0.532	0.530
	HATS	0.562	0.559
Our Method	**CME-HG**	**0.604**	**0.583**

表 2.實驗結果比較

與傳統的回歸模型相比，我們提出的 CME-HG 模型在 Accuracy 和 F1-score 上分別高出約 7% 和 9%。相比於序列式神經網路模型的方法，CME-HG 模型的效果分別提升了約 5% 和 7%。針對使用社群平台言論作為特徵的三種方法，CME-HG 模型相較於 RandForest 與 TSLDA 分別高出約 7% 和 5%，並領先 HAN 模型約 1% 和 2%。而相比於 StockNet 模型（同時結合股價與社群言論特徵，並使用序列式神經網路和注意力機制），CME-HG 模型的準確率和 F1-score 分別高出約 5% 和 3%。此外，針對使用圖神經網路來學習股票間關係的兩種模型，CME-HG 模型比 Chen 等人的研究結果高出約 7% 和 5%，領先 HATS 模型約 2% 和 4%。實驗結果表明，我們的 CME-HG 模型在股票趨勢預測中表現出卓越的效果。

5 分析和討論

我們在此章節分析了不同的特徵，和短、中、長期資料對模型的影響，以及對個案的分析。

5.1 不同特徵的比較

為了更清楚地了解哪些特徵對於股價漲跌的影響，我們進行了單一特徵和兩兩特徵組合的模型訓練，並對比分析了三種特徵對模型表現的重要性。實驗結果如表 3 所示。

平均特徵	價差特徵	情緒特徵	Accuracy	F1-score
o	x	x	0.53	0.606
x	o	x	0.581	0.606
x	x	o	0.523	**0.641**
o	o	x	**0.604**	0.576
o	x	o	0.531	0.622
x	o	o	0.602	0.563
o	o	o	**0.604**	0.583

表 3.不同特徵的效能比較

從這項實驗結果中，我們觀察到兩個特別值得注意的現象。首先，當只使用情緒特徵作為唯一輸入時，模型的 Accuracy 表現最低，僅為 0.523，而 F1-Score 卻達到了最高，為 0.641。為了進一步了解這一現象，我們計算了實驗中的 Recall 和 Precision，分別為 1 和 0.522。這表明模型在預測中所有結果均被判定為「漲」，顯示單單依賴情緒特徵並不能有效地預測股價趨勢，導致分類結果嚴重失衡。其次，透過不同特徵組合的實驗結果，我們發現如果只使用價差特徵（如第二項實驗），或者價差特徵與其他特徵搭配使用（如第四項和第六項實驗），模型的預測表現均較高。這說明價差特徵對於模型的準確預測具有較強的影響力。最後，從第七項實驗可以看到，當三種不同性質的特徵都作為輸入特徵時，模型表現最佳，表明多樣化的特徵融合能夠最大限度地提升模型的預測效果。

5.2 短、中、長期資料的比較

我們嘗試將資料按不同的天數進行取樣，分別是每 1 天（單日）、每 2 天、每 3 天、每 5 天、每 6 天、每 7 天、每 8 天、每 9 天、每 10 天（雙周）和每 20 天（單月）作為資料樣本，以探討模型在學習短期、中期和長期資料時，對預測表現的影響。實驗結果如圖 8 所示，展示了不同時間取樣間隔下模型的預測效果。

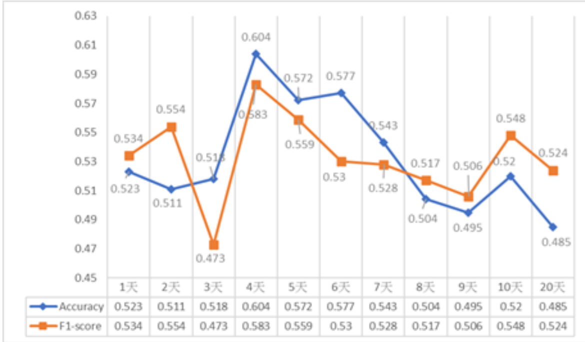


圖 8. 不同時間長度之實驗結果

從實驗結果得知，當模型在訓練中使用 4 天的資料進行學習時，預測表現最佳。隨著取樣天數的增加，模型表現逐漸下降。這是因為隨著時間跨度的增加，中、長期的數據特徵對股價的影響程度降低。此外，社群發文中的情緒隨著時間推移變得模糊，增加了噪聲和干擾，進而導致模型效果下降。

5.3 CASE STUDY

由 CME-HG 模型預測的 85 檔股票中，每檔股票的預測表現不一。我們針對其中預測表現最差的股票——準確率僅為 0.48 的 AAPL（APPLE）進行了深入分析，如圖 9 所示。此圖為 AAPL（APPLE）從 2015 年 10 月 1 日到 2016 年 1 月 1 日的收盤價折線圖。圖中的 x 軸為日期，y 軸為收盤價。紅色點代表模型在當天的預測結果與實際結果相同，即模型預測正確；藍色點代表模型在當天的預測結果與實際結果不同，也就是模型預測錯誤。可以看出，AAPL 股票在 2015 年 11 月 3 日之後，股價從 122 美元快速下跌至約 106 美元，導致模型在這段期間的預測表現較差。經過檢視推文內容後發現，這段期間股價波動較大，發文者對於股價走勢的看法存在多空分歧，這種情緒不一致性導致了模型預測的偏差，從而顯著降低了準確率。這表明在股價波動劇烈且市場情緒不穩定的情況下，模型的預測能力會受到較大影響。

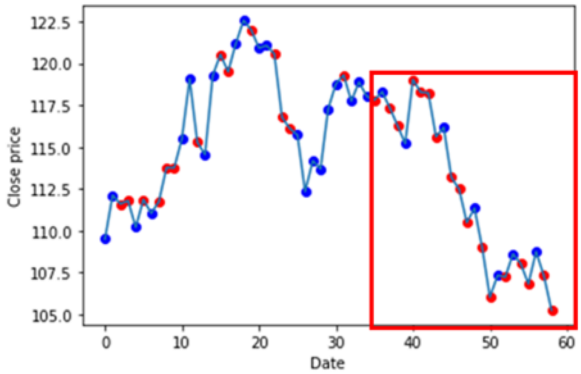


圖 9. AAPL(APPLE)收盤價折線圖(2015/10/01-2016/01/01)

6　結論與未來展望

上市公司股票趨勢受到歷史價格、技術分析、網路言論和股票之間複雜關聯性的影響。本文提出了一種結合歷史股價與社群情緒的多模態整合模型，用於預測股票的漲跌走勢。

該模型利用三種不同的特徵來預測股票的漲跌，包括基於收盤價計算的移動平均線、基於開盤價、最高價、最低價與收盤價計算的股價變動，以及社群言論的情緒分析。移動平均線使用 GRU 提取序列特徵，短期股價變化藉由圖注意力機制獲取價差特徵，社群情緒特徵也通過圖注意力機制得出。我們將三種不同性質的特徵融合為股票的表示形式，最終通過超圖神經網路預測股票的漲跌。

實驗結果顯示，我們所提出的模型在結合多種不同性質的特徵並考慮股票之間的關聯後，Accuracy 達到 0.604。相比先前的方法，我們的模型更有效地捕捉了股票的漲跌趨勢。並且透過實驗可以發現，在這項預測股票趨勢的任務中，價差特徵對模型表現的影響最大，其次是移動平均特徵，最後是情緒特徵。

未來的研究將著重於三個方面的改進。首先，在特徵選取方面，我們計劃引入財經新聞的語義分析作為額外特徵，以進一步提高模型在股票趨勢預測中的準確率。其次，針對情緒特徵，由於本研究中未能精確分類出針對個股的正面與負面情緒，導致情緒特徵中存在噪聲，我們計劃針對這一問題進行更精細的處理，以提升模型的預測準確性。最後，面對模型潛在的過度擬合以及在現實世界中的適用性，我們將考慮使用 regularization techniques 和更多的真實市場分析來進一步改善。

Acknowledgments

The authors would like to thank the supports from the National Science and Technology Council, Taiwan under the grant numbers: NSTC113-2221-E-027-096, and NSTC113-2634-F-027-001-MBK.

References

Amin, M. S., Ayon, E. H., Ghosh, B. P., MD, M. S. C., Bhuiyan, M. S., Jewel, R. M., & Linkon, A. A. (2024). Harmonizing Macro-Financial Factors and Twitter Sentiment Analysis in Forecasting Stock Market Trends. *Journal of Computer Science and Technology Studies*, *6*(1), 58-67.

Asgarov, A. (2023). Predicting Financial Market Trends using Time Series Analysis and Natural Language Processing. *arXiv preprint arXiv:2309.00136*.

Behrendt, S., Schmidt, A. J. J. o. B., & Finance. (2018). The Twitter myth revisited: Intraday investor sentiment, Twitter activity and individual-level stock return volatility. *96*, 355-367.

Ben Cheikh, S., Amiri, H., & Loukil, N. J. I. J. o. S. E. (2024). Social media investors' sentiment as stock market performance predictor. *51*(6), 713-724.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Brown, R. G. (2004). *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation.

Chen, Y., Wei, Z., & Huang, X. (2018). Incorporating corporation relationship via graph convolutional neural networks for stock price prediction. Proceedings of the 27th ACM international conference on information and knowledge management,

Diederik, P. K. J. (2014). Adam: A method for stochastic optimization.

Duan, Y., Wang, L., Zhang, Q., & Li, J. (2022). Factorvae: A probabilistic dynamic factor model based on variational autoencoder for predicting cross-sectional stock returns. Proceedings of the AAAI Conference on Artificial Intelligence,

Fan, J., & Shen, Y. (2024). StockMixer: A Simple Yet Strong MLP-Based Architecture for Stock Price Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence,

Feng, F., He, X., Wang, X., Luo, C., Liu, Y., & Chua, T.-S. (2019). Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems (TOIS)*, *37*(2), 1-30.

Feng, Y., You, H., Zhang, Z., Ji, R., & Gao, Y. (2019). Hypergraph neural networks. Proceedings of the AAAI conference on artificial intelligence, Hu, Z., Liu, W., Bian, J., Liu, X., & Liu, T.-Y. (2018). Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. Proceedings of the eleventh ACM international conference on web search and data mining,

Kim, R., So, C. H., Jeong, M., Lee, S., Kim, J., & Kang, J. (2019a). Hats: A hierarchical graph attention network for stock movement prediction. *arXiv preprint arXiv:1908.07999*.

Kim, R., So, C. H., Jeong, M., Lee, S., Kim, J., & Kang, J. J. a. p. a. (2019b). Hats: A hierarchical graph attention network for stock movement prediction.

Li, T., Liu, Z., Shen, Y., Wang, X., Chen, H., & Huang, S. (2024). MASTER: Market-Guided Stock Transformer for Stock Price Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence,

Li, W., Bao, R., Harimoto, K., Chen, D., Xu, J., & Su, Q. (2021). Modeling the stock relation with graph network for overnight stock movement prediction. Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence,

Lu, W., Li, J., Wang, J., Qin, L. J. N. C., & Applications. (2021). A CNN-BiLSTM-AM method for stock price prediction. *33*(10), 4741-4753.

Mikolov, T. J. a. p. a. (2013). Efficient estimation of word representations in vector space.

Nguyen, T. H., & Shirai, K. (2015). Topic modeling based sentiment analysis on social media for stock market prediction. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),

Nuñez-Mora, J. A., Mendoza-Urdiales, R. A. J. S. N. A., & Mining. (2023). Social sentiment and impact in US equity market: an automated approach. *13*(1), 111.

Ranjith, J. (2024). Enhancing Stock Market Trend Prediction Using Explainable Artificial Intelligence and Multi-source Data. *Fusion: Practice and Applications*, *16*(2), 178-178-189.

Sawhney, R., Agarwal, S., Wadhwa, A., Derr, T., & Shah, R. R. (2021). Stock selection via spatiotemporal hypergraph attention network: A learning to rank approach. Proceedings of the AAAI Conference on Artificial Intelligence,

Sawhney, R., Agarwal, S., Wadhwa, A., & Shah, R. (2020). Deep attentive learning for stock movement prediction from social media text and company correlations. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),

Selvin, S., Vinayakumar, R., Gopalakrishnan, E., Menon, V. K., & Soman, K. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. 2017 international conference on advances in computing, communications and informatics (icacci),

Tabari, N., Biswas, P., Praneeth, B., Seyeditabari, A., Hadzikadic, M., & Zadrozny, W. (2018). Causality analysis of Twitter sentiments and stock market returns. Proceedings of the first workshop on economics and natural language processing,

Wang, S., Bai, Y., Ji, T., Fu, K., Wang, L., & Lu, C.-T. (2023). Stock Movement and Volatility Prediction from Tweets, Macroeconomic Factors and Historical Prices. 2023 IEEE International Conference on Big Data (BigData),

Xu, Y., & Cohen, S. B. (2018). Stock movement prediction from tweets and historical prices. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),

Yang, H., Zeng, B., Xu, M., & Wang, T. (2021). Back to reality: Leveraging pattern-driven modeling to enable affordable sentiment dependency learning. *arXiv preprint arXiv:2110.08604*.

Zhang, L., Aggarwal, C., & Qi, G.-J. (2017). Stock price prediction via discovering multi-frequency trading patterns. Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining,

殘差模組結合擠壓與激發注意力機制改進少樣本道路警報偵測模型
Residual Modules Combined with Squeeze-and-Excitation Attention Mechanism for Improving Few-Shot Road Alert Detection Model

郭子豪 **Tzu-Hao Kuo**
國立中山大學資訊工程學系
National Sun Yat-sen University
Department of Computer Science and Engineering
m113040053@student.nsysu.edu.tw
鄭羽涵 **Yu-Han Cheng**
中華電信研究院前瞻科技研究所
Advanced Technology Laboratory, Chunghwa Telecom Laboratories
henacheng@cht.com.tw
陳嘉平 **Chia-Ping Chen,** 陳品錞 **Ping-Chun Chen**
國立中山大學資訊工程學系
National Sun Yat-sen University
Department of Computer Science and Engineering
cpchen@cse.nsysu.edu.tw
pipichun17@gmail.com
呂仲理 **Chung-Li Lu,** 詹博丞 **Bo-Cheng Chan,** 莊向峰 **Hsiang-Feng Chuang,**
陳威妤 **Wei-Yu Chen**
中華電信研究院前瞻科技研究所
Advanced Technology Laboratory, Chunghwa Telecom Laboratories
{chungli, cbc, gotop, weiweichen}@cht.com.tw

摘要

本文提出了一個少樣本分類模型 SE-RCNN。該模型使用殘差模組來提升特徵擷取能力，並利用 GELU 激勵函數改善 ReLU 資訊丟失的問題，同時通過擠壓與激發注意力機制強調關鍵特徵。在 5-way 1-shot 的少樣本情境下，該模型在 ESC-50 資料集上的準確率從 76.2% 提升至 82.1%。接著，我們以此模型爲雛型，在 4-way 的情境下利用自行蒐集的道路警報資料集進行調適。我們的模型在每個類別提供 15 個樣本的情形下各類別的 F1-score 均不低於 0.8。最後，我們以片段級預測的方式實做一個道路警報偵測模型。

Abstract

This paper proposes a few-shot classification model called SE-RCNN. The model uses residual modules to enhance feature extraction capabilities, GELU activation functions to mitigate information loss from ReLU, and Squeeze-and-Excitation attention mechanisms to emphasize key features. In a 5-way 1-shot few-shot learning scenario, the model's accuracy on the ESC-50 dataset improved from 76.2% to 82.1%. Subsequently, we used this model as a prototype and adjusted it using the self-collected road alert dataset in a 4-way scenario. Under the condition of providing 15 samples for each category, our model achieved an F1-score of no less than 0.8 for all categories. Finally, we implemented a road alert detection model using a segment-level prediction approach.

關鍵字：聲音事件偵測、少樣本學習、注意力機制

Keywords: sound event detection ,few-shot learning, attention mechanism

1 緒論

聲音事件偵測 (Sound Event Detection, SED) 是一種自動從聲音訊號中識別和分類各種聲音事件的技術。這些聲音事件涵蓋環境音效（如雷聲、狗吠）、人爲活動（如講話、拍手）以及機械聲（如救護車警笛、汽車引擎聲）等。SED 技術在智慧城市、監控系統、醫療監測和人機交互等領域具有廣泛的應用。例如，在工業環境中，異常聲音檢測系統能夠及早識別需要維護的潛在故障設備 (Dohi et al., 2022)。在家庭環境中，偵測系統可以即時偵測破窗聲

或火災警報等事件，從而實現即時警報和緊急應變。然而，在實際應用中，資料蒐集經常面臨挑戰。由於聲音事件的種類繁多，蒐集大量資料集涉及龐大的人力成本，某些特殊聲音事件甚至難以進行大規模蒐集。因此，我們嘗試結合少樣本學習 (Few-Shot Learning) 與聲音事件偵測，以解決這一問題。

少樣本學習是機器學習的一個特殊分支，旨在使模型在面對新任務時，能夠依賴少量資料進行訓練，並展現出良好的泛化能力。這一概念在資料蒐集困難且類型繁多的聲音事件偵測領域具有巨大的應用潛力。然而，目前的少樣本學習研究主要集中在圖像領域。例如，Vinyals et al. (2016) 與 Song et al. (2020) 分別使用圖像資料集進行訓練並應用於圖像識別任務。此外，大多數研究主要著眼於演算法性能的提升或公開資料集準確率的比較，卻忽略了少樣本學習在實際應用中的便利性和優勢。

因此，本論文提出了一個道路警報偵測模型，該系統能夠識別道路上需禮讓的緊急車輛聲音，如救護車、警車和消防車。在當今社會，隨著車輛隔音技術的進步，佩戴耳機的行人和騎士在道路上的比例逐漸上升，這使得該系統能夠有效提醒用路人對緊急事件車輛進行禮讓，從而提升整體的交通安全環境。我們的模型基於 MetaAudio (Heggan et al., 2022) 作爲核心分類模型，並通過引入殘差網路結構 (He et al., 2016)、高斯誤差線性單元 (Hendrycks and Gimpel, 2016) 以及擠壓與激發注意力機制 (Hu et al., 2018)，來增強基礎模型的特徵擷取能力。隨後，我們利用自行蒐集的道路警報資料集對模型進行微調。最終，我們的模型可以進行片段級別的預測。

本文的後續章節安排如下：第二章：研究方法，將說明模型架構與訓練方式；第三章：實驗設置，將介紹實驗的相關設置與資料集；第四章：實驗結果，將比較不同模型之間的差異並說明相關實驗數據；第五章：結論。

2 研究方法

在這個章節中，我們將詳細說明本次實驗所使用的各種方法，包括少樣本學習的方法、所使用的骨幹網路架構細節，以及使用的注意力機制。實驗的核心基礎主要基於 MetaAudio (Heggan et al., 2022)，並對其骨幹網路架構進行了改良。

2.1 模型架構

此部分將說明基礎模型與我們改良的模型，並進一步說明高斯誤差線性單元以及擠壓與激發注意力機制的相關內容。

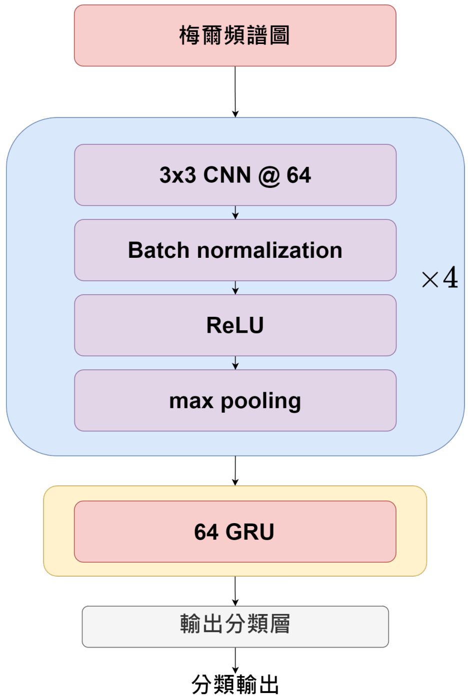


圖 1: CRNN 模型: 輸入頻譜圖首先經過四個堆疊的卷積模組，之後通過單層的 GRU，最後由輸出分類層輸出結果。

2.1.1 基礎模型

MetaAudio 使用兩種不同的基礎模型，分別是卷積神經網路 (Convolutional Neural Network, CNN) 和卷積遞歸神經網路 (Convolution-Recurrent Neural Network, CRNN)。CRNN 在 CNN 的基礎上進一步擴展，加入了一層遞歸神經網路（Recurrent Neural Network, RNN）。CNN 架構由四個卷積模組組成，每個模組均使用 3×3 的卷積層，隨後是批正規化層（Batch Normalization）、線性整流函數（Rectified Linear Unit, ReLU (Agarap, 2018) 以及最大池化層（Max Pooling）。其中，最大池化層的取值範圍皆爲 2×2。在 CRNN 中，所使用的 RNN 架構基於閘門遞歸單元（Gated Recurrent Unit, GRU） (Chung et al., 2014)。GRU 通過其門控機制，有效地捕捉序列資料中的時間依賴性，這使得 CRNN 能夠在處理時序音訊資料時更好地保留和利用上下文資訊。最後，無論是 CNN 還是 CRNN，兩者都連接至一個分類層。該分類層由一個具有 30% 丟棄率的丟棄層（Dropout）、一層批正規化層和一層全連接層（Fully Connected Layer）組成。CRNN 整體架構圖如圖 1。

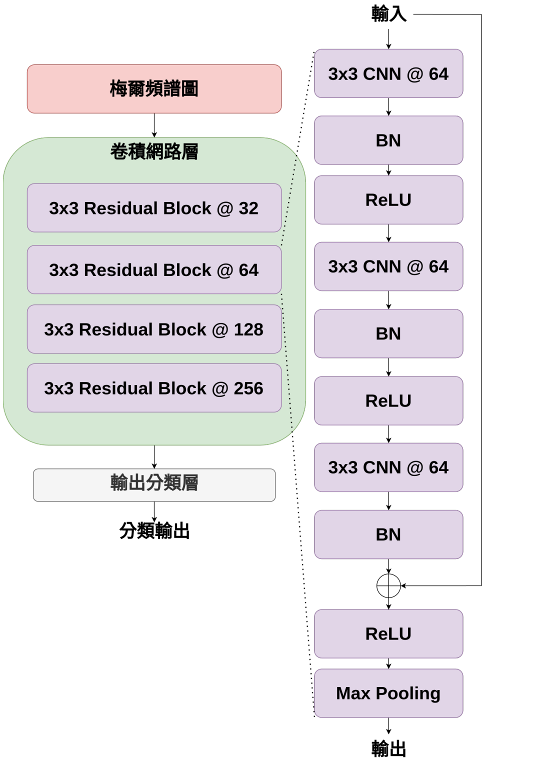


圖 2: RCNN 模型: 輸入頻譜圖首先經過四個堆疊的殘差模組，四個殘差模組的通道數依序爲 32、64、128 與 256，最後由輸出分類層輸出結果。

2.1.2 殘差卷積神經網路模型

我們選擇以 CNN 模型作爲改進的基礎，主要的改動想法是增加模型的整體深度，以便使模型能夠學習到更爲複雜的特徵表示。透過增加深度，模型可以捕捉到更細緻的資料特徵，從而提高其表現。爲了實現這一目標，我們引入了捷徑連接（shortcut connection） (He et al., 2016)，並結合多個卷積層構建殘差模組 (residual module)。這種殘差結構使得資訊在模型中可以有效地流動，並減輕了隨著層數增加而可能出現的梯度消失問題。具體而言，捷徑連接允許原始輸入直接傳遞到更深的層，從而促進了更快的訓練和更好的收斂效果。我們將修改後的模型命爲 RCNN。RCNN 由四層殘差模組與一層全連接層組成。殘差模組中包含兩個分支，一個分支是恆等映射函數，另一個分支是多層結構。該結構依序由一個卷積層、批正規化層、ReLU 激勵函數、卷積層、批正規化層、ReLU 激勵函數、卷積層與批正規化層組成。這兩個分支的輸出相加後，再經過 ReLU 激勵函數和最大池化層進行處理。每個殘差模組的通道數分別爲 32、64、128 和 256。CRNN 整體架構圖如圖 2。

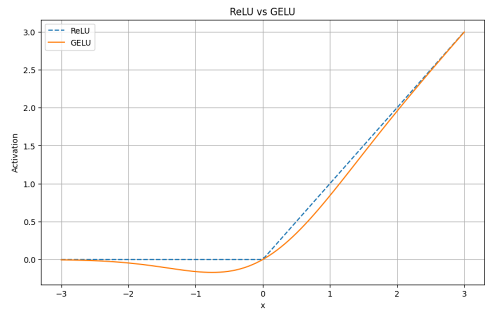


圖 3: 高斯誤差線性單元與線性整流函數圖：圖中橘色實線爲高斯誤差線性單元，藍色虛線爲線性整流函數。

2.1.3 高斯誤差線性單元

高斯誤差線性單元（Gaussian Error Linear Units, GELU (Hendrycks and Gimpel, 2016)）是一種激勵函數。與常見的 ReLU 激勵函數相比，GELU 在輸入值接近 0 的區域具有平滑的轉換，這有助於避免梯度的突然變化，從而提高模型訓練過程的穩定性。公式 1 爲 GELU 的數學表達式：

$$\mathrm{GELU}(x) = x \cdot \Phi(x) \tag{1}$$

$\Phi(x)$ 代表高斯分佈的累積分佈函數，可由公式 2 表示：

$$\Phi(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \tag{2}$$

erf 表示爲誤差函數 (error function)。爲了便於計算，GELU 通常會採用近似公式來簡化計算。公式 3 是常用的 GELU 近似公式：

$$\begin{aligned} \mathrm{GELU}(x) &\approx 0.5x(1+T) \\ T &= \tanh\left[\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right] \end{aligned} \tag{3}$$

圖 3 爲 GELU 激勵函數與 ReLU 激勵函數的比較圖。

2.1.4 擠壓與激發注意力機制

此外，我們還嘗試引入擠壓與激發注意力機制（Squeeze-and-Excitation attention, SE attention） (Hu et al., 2018)，SE attention 的結構如圖 4。注意力機制能夠幫助模型自動聚焦於重要的特徵，進一步增強了模型的表示能力。SE attention 它通過擠壓特徵圖的通道資訊，生成一個權重向量，然後根據這些權重自適應地調整各通道的特徵強度。具體而言，

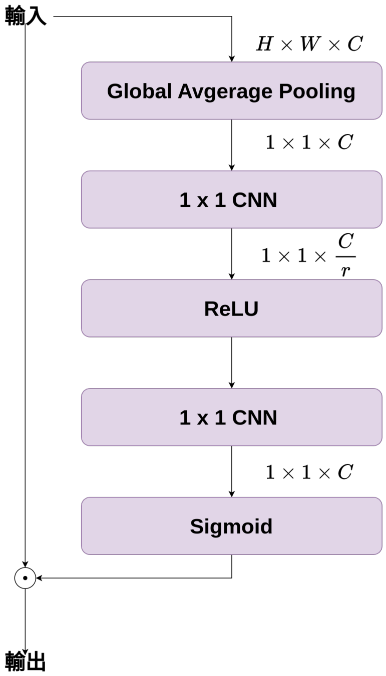


圖 4: SE attention 結構圖：C 代表特徵通道數，r 代表擠壓率，⊙ 代表元素對應相乘。輸入先經由全局池化層將每個通道的空間維度進行平均，產生一個 $1 \times 1 \times$ C 的向量，之後經由兩個卷積操作擠壓與激發向量。最後經過乙狀函數 (Sigmoid) 產生通道權重向量。

SE attention 包括兩個步驟：首先對特徵圖進行全局平均池化以獲取通道向量，然後通過全連接層或 1×1 卷積生成權重向量。這種方式能夠強調重要特徵，抑制不重要特徵，提升模型表現。

2.2 少樣本學習

少樣本學習的訓練過程與一般深度學習有所不同，其目標是能夠快速適應「新」任務，因此在少樣本學習中，每一筆訓練資料的單位是一個「任務」。任務會被劃分爲支持集（support set）與查詢集（query set），其中支持集與查詢集彼此互斥。支持集的功能類似於訓練集，而查詢集則接近於測試集的概念。任務通常會設定一個情境，說明需要分類的類別數量以及每個類別提供的帶標籤資料數量，並以 N-way K-shot 來表示。N-way 說明任務中有 N 個不同的目標類別，K-shot 則表示每個目標類別有 K 個樣本。例如，在一個 5-way 5-shot 的任務中，支持集包含 5 個類別，每個類別有 5 筆樣本，而查詢集通常包含 5 個類別，每個類別有 1 個樣本。在少樣本學習的訓練過程中，模型會接觸到很多個任務，每個任務都是從訓練集中抽樣。通過這樣的方式，模型不僅能夠學習到如何快速適應新任務，還能夠捕捉到不同任務之間的關聯性。這種任務間的相互作用有助於模型更好地理解特徵的共享性，從而提升其在新任務上的性能。

2.2.1 元曲率

元曲率 (Meta-Curvature) (Park and Oliva, 2019) 是我們主要使用的少樣本學習方法，旨在訓練時學習通用的模型參數與更新的曲率，以提高模型對於新任務的是適應能力。爲了實現這一目標，訓練過程會將「任務」是爲訓練資料的單位，模型反覆在不同的任務上進行訓練，以便訓練後的模型能夠使用少量的樣本或訓練迭代適應新任務。

首先，元曲率定義了三個元曲率矩陣，$M_o \in \mathbb{R}^{C_{out} \times C_{out}}$、$M_i \in \mathbb{R}^{C_{in} \times C_{in}}$ 和 $M_f \in \mathbb{R}^{d \times d}$，$C_{out}$、$C_{in}$ 和 d 分別表示輸出通道數、輸入通道數和濾波器大小。在卷積層中，d 表示高度 $\times$ 寬度，而在全連接層中，d 表示 1。元曲率的函數定義如公式 4 所述：

$$MC(\mathrm{G}) = G \times M_f \times M_i \times M_o \tag{4}$$

G 代表損失函數的梯度，且 M_f、M_i 與 M_o 在初始化時爲單位矩陣。

假設 f_ϕ 代表一個參數爲 ϕ 的模型。從訓練資料集隨機抽樣 k 個任務，k 是一個可自訂的超參數，以 $\mathcal{T}_i$ 表示這 k 個任務中的第 i 個任務。f_ϕ 以任務的支持集作爲訓練集訓練參數，更新後的參數爲 θ_i。如公式 5 所述：

$$\theta_i = \phi - \alpha MC(\nabla \mathcal{L}_{\mathcal{T}_i}(\phi)) \tag{5}$$

θ_i 是經過 $\mathcal{T}_i$ 支持集訓練後的參數，而 α 爲學習率，並將此參數更新稱作內循環 (inner loop)。之後，每個任務使用各自的查詢集再次計算損失，根據這些損失更新模型參數 ϕ 與矩陣參數 M_f、M_i 與 M_o。如公式 6 所述：

$$\begin{aligned}
\phi &\leftarrow \mathrm{ADAM}(\phi, \beta, \nabla_\phi \sum_{\mathcal{T}_i} \mathcal{L}_{\mathcal{T}_i}(\theta_i)) \\
M_o &\leftarrow \mathrm{ADAM}(M_o, \beta, \nabla_{M_o} \sum_{\mathcal{T}_i} \mathcal{L}_{\mathcal{T}_i}(\theta_i)) \\
M_i &\leftarrow \mathrm{ADAM}(M_i, \beta, \nabla_{M_i} \sum_{\mathcal{T}_i} \mathcal{L}_{\mathcal{T}_i}(\theta_i)) \\
M_f &\leftarrow \mathrm{ADAM}(M_f, \beta, \nabla_{M_f} \sum_{\mathcal{T}_i} \mathcal{L}_{\mathcal{T}_i}(\theta_i))
\end{aligned} \tag{6}$$

β 代表學習率，ADAM 爲我們使用的優化器 (optimizer)，並將此參數更新稱作外循環 (outer loop)。以上爲元曲率的一次迭代訓練，流程圖如圖 5。

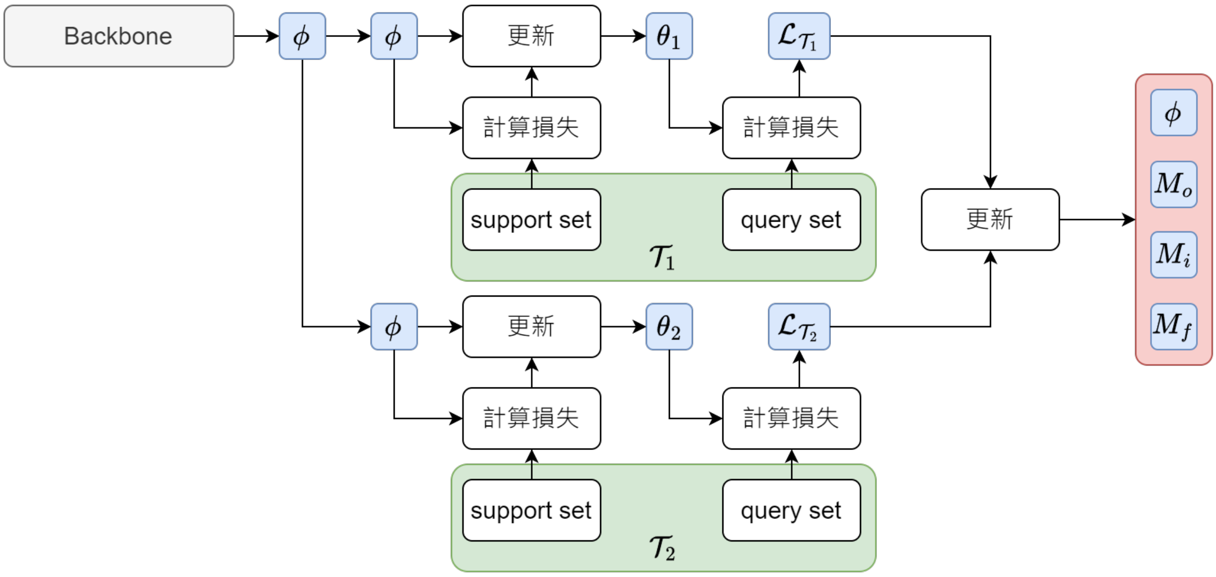


圖 5: Meta-Curvature 訓練的示意圖：此圖爲 Meta-Curvature 訓練的舉例說明。假設任務數量設定爲 2，模型的參數 ϕ 將被複製爲兩份。這兩份參數分別使用 $\mathcal{T}_1$ 與 $\mathcal{T}_2$ 進行參數更新。最後，通過在查詢集（query set）的損失總合更新參數 ϕ、M_o、M_i 與 M_f。

3 實驗設置

本節將描述本論文實驗中所採用的相關設置，包括模型使用的訓練集與測試集、聲音特徵的前處理方法、訓練過程中的學習率設置及各項超參數的設定。此外，還將說明用於評估模型表現的效能指標。

3.1 資料集

我們選擇使用 ESC-50 資料集 (Piczak, 2015) 作爲訓練與測試雛型模型的資料集。ESC-50 是一個環境聲音分類資料集，包含 2,000 個標註音檔，每個音檔長度爲 5 秒，採樣率爲 44.1kHz，檔案格式爲.wav。這些音檔來自 50 個不同的類別，每個類別各包含 40 筆資料。由於少樣本訓練的特殊性，我們將資料集依據類別進行分割：35 個類別用作訓練集，5 個類別作爲驗證集，其餘 10 個類別則用於測試。此分割方式參考了 Chou et al. (2019)。我們自行蒐集了一個道路警報資料集，該資料集包含四種類別的聲音，分別爲救護車聲、警車聲、消防車聲以及道路雜音。道路雜音類別擁有 300 筆音檔，而其他每種類別則各有 100 筆音檔。所有音檔均爲 5 秒長，且採樣率爲 16kHz。此資料集將用作訓練與測試道路警報模型的資料集。

3.2 聲音特徵前處理方法

首先，我們將所有音檔重新採樣至 16kHz，接著將音訊從波形訊號（waveform）轉換爲梅爾頻譜圖（mel-spectrogram），最後取對數作爲模型的輸入。在頻譜圖的設定中，使用 128 個梅爾濾波器 (Melfilter bank)，傅立葉轉換的窗長 (Window size) 設爲 1,024，跳躍長度 (Hop length) 爲 512。

3.3 參數設定

本論文的所有實驗數據均使用相同的訓練設定。每個模型訓練 200 個 epoch，並採用 Adam 優化器 (optimizer) (Kingma and Ba, 2014)，內循環學習率 α 設爲 0.4，外循環學習率 β 設爲 0.001，整個過程中學習率均保持固定，不使用自適應技術。在驗證階段，我們從驗證集中隨機抽取 200 個任務進行準確率測試；在測試階段，則從測試集中隨機抽取 5,000 個任務進行評估。任務由支持集與查詢集組成，模型先以支持集調適模型，然後在查詢集上進行測試以計算準確率。最終呈現之準確率爲所有任務的平均準確率。

3.4 評估準則

我們主要以準確率 (Accuracy) 作爲雛型模型的評估準則，準確率是正確預測的樣本數與總樣本數的比率。由於道路警報系統模型的測試集存在類別不平衡的現象，因此在道路警報系統模型的部份則使用 F1-Score 作爲評估準則。F1-Score 是精確率 (Precision) 和召回率 (Recall) 的調和平均數，公式 7 8 9 別分呈現精確率、召回率與 F1-Score 的計算方式：

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

模型架構	Accuracy(%)
CNN	71.2
CRNN	76.2
RCNN	80.1

表 1: RCNN 與基礎模型準確率比較：展示 CNN、CRNN 和 RCNN 模型在 ESC-50 測試集上進行 5,000 個任務抽樣後的平均準確率結果。

模型架構	Accuracy(%)
RCNN	80.1
RCNN(GELU)	80.8
SE-RCNN(GELU)	82.1

表 2: RCNN 改動準確率比較：呈現 RCNN 更換激勵函數與引入注意力機制後，於 ESC-50 測試集上進行 5,000 個任務抽樣後的平均準確率結果。

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$\text{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

TP 代表正確預測爲正類的樣本數，FP 代表錯誤預測爲正類的樣本數，FN 代表錯誤預測爲負類的樣本數。

4 實驗結果

4.1 模型比較

表 1顯示我們所改進的 RCNN 模型相較於基礎模型展示了顯著的進步，這一進步主要體現在殘差模組的設計上。殘差模組通過堆疊多層卷積結構，使模型能夠有效地學習和捕捉更深層次的特徵關係。具體來說，這些堆疊的卷積層增強了模型的表達能力，使其能夠更好地捕捉數據中的複雜模式和細微變化。同時，RCNN 模型還對通道數進行了擴展，進一步增強了模型處理不同特徵的能力。此外，殘差模組引入了捷徑連接，不僅能減少梯度消失問題，還能提升深層神經網路的訓練穩定性和效率。捷徑連接通過直接將輸入傳遞到後面的層，形成一條捷徑路徑，使得梯度可以更順暢地反向傳播，從而有效避免深層網路常見的梯度消失問題。

少樣本情境	Accuracy(%)	記憶體佔用
1-shot	82.1	4.9GB
3-shot	94.4	10.3GB
5-shot	96.3	*9.4GB

表 3: shot 數與準確率關係：呈現在不同少樣本情境下訓練 SE-RCNN 模型後，於 ESC-50 測試集上進行 5,000 個任務抽樣的平均準確率結果與記憶體佔用情形。標記 * 表示因超出硬體極限而進行的參數調整後的記憶體佔用情形。

4.2 RCNN 改動比較

表 2 顯示更換激勵函數爲 GELU 與引入注意力機制對於 RCNN 準確率提昇的有效性。替換激勵函數後準確率提 0.7 %，這是因爲相比於 ReLU 在輸入小於零時會導致神經元關閉，可能會丟失一些有用的資訊，而 GELU 通過引入高斯分佈，使激勵函數能夠對輸入進行更加平滑的處理，從而保留更多的細節。引入 SE attention 後，模型的準確率又提升了 1.3 %，因爲 SE attention 的引入使模型能夠自適應地重新調整特徵的權重，從而增強對重要特徵的關注，抑制不重要的特徵。這種動態調整特徵權重的方法，有助於模型更加精確地捕捉關鍵資訊，從而提升整體性能。

4.3 K-shot 情境分析

此實驗主要目的是爲了解 K-shot 對於預測的影響趨勢，我們固定 5-way 進行訓練，分析不同 shot 數對準確率與硬體需求的影響。表 3 呈現在 5-way 不同 shot 數情境下使用 SE-RCNN 模型的訓練結果。隨著 shot 數量的增加，模型能夠獲取更豐富的資料和多樣的特徵，這使得模型更容易擷取各類別的共同特徵，從而提高模型的泛化能力。因此，表格中 5-shot 的訓練結果顯示出最佳的準確率。然而，隨著 shot 數量的增加，訓練對硬體的需求也相應上升。以本次訓練爲例，5-way 5-shot 訓練已達到我們訓練設備 NVIDIA 1080 Ti 顯示卡的記憶體極限，因此我們有對 5-way 5-shot 的訓練參數進行調整。後續道路警報系統實驗皆以 5-shot 情境下的 SE-RCNN 作爲雛型模型。

4.4 調適模型

少樣本訓練到模型應用的過程分爲兩個階段。第一階段是利用一定量的樣本來訓練一個雛型模型。例如，本文採用 ESC-50 資料集，通過元曲率方法訓練出一個雛型模型。第二階段則是對雛型模型進行調適，使其適應特定的目

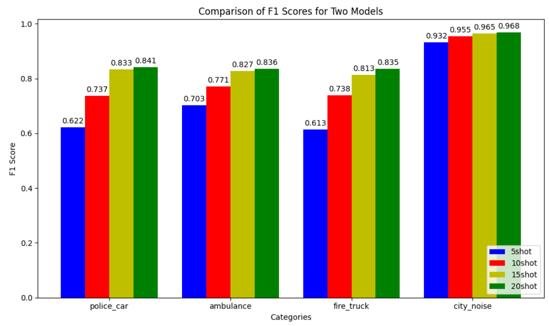


圖 6: F1-Score 比較圖：此表展示 SE-RCNN 模型在不同 shot 數下，對道路警報資料進行支持集抽樣以調適模型，並將剩餘樣本作爲測試集，重複 100 次後得到的各類別平均 F1-Score 結果。

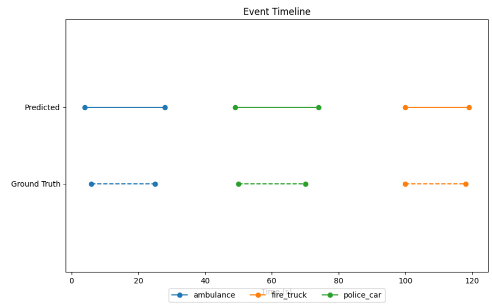


圖 7: 音檔預測示意圖：此圖不同顏色的線段代表不同事件，虛線爲事件時間的眞實標籤，實線爲事件時間之預測。

標任務。在本文中，我們將雛型模型調適至道路警報偵測任務。在第二階段，我們以 4-way 的形式對模型進行調適。道路警報模型是基於一個 4-way 的 SE-RCNN 雛型模型進行調適而成。我們使用道路警報資料集作爲調適階段的訓練集。模型的測試方式是通過從訓練集隨機抽樣生成支持集來調適模型，然後使用訓練集的剩餘樣本作爲測試集來評估模型的效能。這一過程重複進行 100 次，以確保結果的穩定性和可靠性，最終呈現各類別的平均 F1-score。圖 6 呈現在調適階段以不同 shot 數訓練的結果。由圖可以觀察到當 shot 越高效能越好，其中以 5-shot 到 10-shot 的提高效果最爲明顯。此外，道路警報模型的 5-shot 結果與 ESC-50 5-shot 結果之間存在一定的差距，我們認爲這主要與資料集的特性有關。爲了更貼近實際應用情況，我們的資料集並未對數據進行額外處理，這使得聲音更容易受到設備和外部環境的影響。相較於 ESC-50 資料集中的聲音，後者的音質較爲乾淨和清晰。

4.5 道路警報偵測系統

我們以章節 4.4 的模型作爲基礎實做道路警報偵測。在偵測系統中，對於每一筆輸入的音檔，系統內部設定了一個固定長度的偵測窗，該偵測窗的長度爲 5 秒。偵測窗以每 1 秒的步伐位移，經過整份音檔完成偵測。輸出的標籤會經過後處理，系統將連續相同類別的窗框區間定義爲事件發生區間，只有區間大於 7 秒的事件會被留下，過小的區間事件將被刪除以穩定輸出。圖 7 爲系統對於音檔預測的示意圖。從圖中可以看出，該系統已經能夠大致辨識出聲音事件，但由於採用片段級預測作爲輸出，若需進行精確定位，其精度仍然不足。因此，這將是我們接下來研究中需要努力改進的部分。

5 結論

本文以 MetaAudio 作爲基礎，嘗試引入殘差模組、GELU 激勵函數與 SE attention，以進一步提升基礎模型的表現。這些改進措施旨在加強模型對聲音特徵的捕捉和辨識能力。經過這些改進，使用元曲率少樣本學習方法，模型的分類準確率從 76.2% 提升至 82.1%。此外，我們以上述模型爲雛型，利用自行蒐集的道路警報資料集調適出一個道路警報偵測模型，並以此建構了一個道路警報偵測系統。然而，該系統仍存在偵測單位過大的問題，這將是我們後續努力的重點目標。

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375.*

Szu-Yu Chou, Kai-Hsiang Cheng, Jyh-Shing Roger Jang, and Yi-Hsuan Yang. 2019. Learning to match transient sound events using attentional similarity for few-shot sound recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 26–30. IEEE.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555.*

Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Takashi Endo, Masaaki Yamamoto, and Yohei Kawaguchi. 2022. Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques. *arXiv preprint arXiv:2206.05876.*

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Calum Heggan, Sam Budgett, Timothy Hospedales, and Mehrdad Yaghoobi. 2022. Metaaudio: A few-shot audio classification benchmark. In *International Conference on Artificial Neural Networks*, pages 219–230. Springer.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Eunbyung Park and Junier B Oliva. 2019. Meta-curvature. *Advances in neural information processing systems*, 32.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

Gege Song, Zhulin Tao, Xianglin Huang, Gang Cao, Wei Liu, and Lifang Yang. 2020. Hybrid attention-based prototypical network for unfamiliar restaurant food image few-shot recognition. *Ieee Access*, 8:14893–14900.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

利用遮罩增強的語言模型校正技術提升中文醫療語音的自動語音識別效果 (Enhancing Automatic Speech Recognition for Chinese Medical Speech Using Masking-Enhanced Language Model Correction)

林璟芸 Jing-Yun Lin; 李旭清 Xu-Qing Li; 鍾聖倫 Sheng-Luen Chung
國立臺灣科技大學電機工程學系
Electrical Engineering Department
National Taiwan University of Science and Technology
Taipei, Taiwan
slchung@mail.ntust.edu.tw; fdsa3304@gmail.com; stanley890314@gmail.com

摘要

本研究優化了一個針對中文醫療語音的自動語音識別系統 (ASR)。現有的 ASR 系統如 Google 的 STT 和 OpenAI 的 Whisper，雖然在一般語音識別中表現良好，但在醫療領域識別準確度不足。為此，本研究提出「遮罩增強語言模型校正技術」，通過遮罩機制對 ASR 輸出進行校正，以提升識別精度。方法包括微調 CKIP BERT 模型構建 MedCKIPBERT，並提出同音異字替換策略。結果顯示，Google STT 的關鍵詞錯誤率從 15.37%降至 13.64%，Whisper 則從 20.82%降至 15.80%。該方法在醫療語音識別中展現潛在價值。

Abstract

This study optimized an Automatic Speech Recognition (ASR) system for Chinese medical speech. Existing ASR systems like Google's STT and OpenAI's Whisper perform well in general speech recognition but lack accuracy in the medical field. To address this, we proposed a "mask-enhanced language model correction technique" that uses masking to correct ASR outputs, thereby improving recognition accuracy. The methods include fine-tuning the CKIP BERT model to build MedCKIPBERT and implementing a homophone substitution strategy. Results showed that the keyword error rate (KER) of Google STT decreased from 15.37% to 13.64%, and Whisper's KER dropped from 20.82% to 15.80%. This approach demonstrates potential value in medical speech recognition.

關鍵字：自動語音識別 (ASR)、中文醫療語音、醫療中文 BERT、遮罩增強的語言模型

Keywords: Speech Recognition (ASR), Chinese Medical Speech, Medical Chinese BERT, Masking-Enhanced Language Model

1 簡介

1.1 醫療語音辨識

隨著 ASR 技術的進步，其在醫療領域的應用日益增多，尤其在自動生成醫護交班病歷和病歷數位化方面展現出顯著效益。

自動生成醫護交班病歷能節省醫護人員手動記錄病人狀況的時間，降低錯誤率，並促進病歷的數位化與標準化。Chung et al. (2021) 利用中文醫療語音資料庫 psChiMeS-14 和 Joint CTC、Transformer 等模型，提升 ASR 系統表現，並建立「ChiMeS+」醫療語音辨識系統。此外，Enarvi et al. (2020) 探討了通過 ASR 與神經網絡摘要生成醫療報告，減少診斷過程中的繁瑣記錄工作。

儘管語音辨識技術在日常生活中已普及，但在醫學、金融、法律等專業領域仍面臨挑戰。首先，專業語料庫蒐集困難且昂貴，需涵蓋多種口音與情境並精確標記。其次，專業術語的發音和拼寫差異增加辨識難度，不同地區用詞差異也影響準確性。此外，語音辨識系統在未見過的醫療科別中可能無法正確識別新關鍵字。中英文交錯時，華人發音的音節化問題也會導致識別錯誤，如將“hemovac”識別為“黑魔法”。這些因素使專業領域的語音辨識應用充滿挑戰。

1.2 貢獻

本論文提出利用 BERT 輔助的遮罩語言模型 (MLM) 校正 ASR 輸出的方法，無需重新訓練 ASR 或蒐集大量專業領域語料庫，只需串接微調過的 BERT 即可顯著提升專業術語的辨識精度。此外，論文還構建了醫學語文模型 MedCKIPBERT，透過微調 CKIP BERT (https://huggingface.co/ckiplab/bert-base-chinese)，支持醫療文本分類、命名實體識別 (NER)、臨床決策支援等醫療 NLP 應用，助力醫療數位化和智能化。

1.3 本文架構

本論文各節內容如下：第二節「串接 MLM 校正架構」回顧文獻，涵蓋 ASR 模型及其糾錯技術，介紹了 Medical Bert 模型 MedCKIPBERT，並討論 CER 與 KER 績效指標。第三節「MLM 的方法」探討了 Mask-MLM 和近似音-MLM 的 ASR 校正技術。第四節「實驗與結果」展示了測試集選擇及其結果，分析了 Google STT 與 Whisper Medium 模型結合 BERT 在醫療語音中的表現，並進行效果對比。最後，第五節「結論」總結研究成果與應用意義。

2 串接 MLM 校正架構

2.1 文獻審閱

本段文獻審閱回顧了 ASR 技術的發展，從傳統架構到端對端模型的成功應用，並探討了領先系統如 Google STT 和 OpenAI Whisper，這些系統在多語言識別中展現出色能力。最後，分析了 ASR 錯誤校正技術，特別是 BERT 模型及其變體在醫療文本中的應用與提升準確性的潛力。

ASR 技術：傳統的 ASR 系統由三個主要模組組成：聲學模型、發音模型和語言模型。聲學模型 (Acoustic model) 負責將語音信號轉換為音素 (phoneme) 的概率分布；發音模型 (Pronunciation model) 則將這些音素轉換為詞或短語，依賴發音詞典來完成此過程；語言模型 (Language model) 根據語料庫中的詞語頻率和上下文，提供詞語出現的概率，使得語句在語法和語義上更合理。這三個模型各自獨立訓練，且需大量的訓練資料支持，然而，若前一個模型出現錯誤，會影響後續模型的準確性，導致整體識別表現不佳。這種架構的限制使得傳統 ASR 系統在應對複雜語音環境時的表現有所不足。

端對端 ASR 模型：隨著深度學習技術的發展，端對端 (End-to-End, E2E) ASR 模型逐漸成為研究的熱點。這些模型通常使用一個單一的深度神經網絡來直接從語音信號預測出文字輸出，從而簡化傳統 ASR 的多組件結構。主要有三種方法：連結時序分類 (CTC) (Amodei et al., 2016) 透過刪除重複字和空白標籤來自動對齊輸出與輸入，而注意力機制 (Attention-based models) (Chan et al., 2016)允許模型動態"關注"輸入序列的不同部分，根據隱藏層狀態和編碼器的時序資訊決定輸出。自注意力機制 (Self-attention) (Vaswani et al., 2017) 則專注於全局依賴關係，使 Transformer ASR 能平行處理時序資料，提升訓練速度和效率。

領先的 ASR 模型： 領先的 ASR 模型包括 Google STT 和 Whisper。Google STT (Zhang et al., 2023) 使用 1200 萬小時音訊數據及 Conformer 架構，擁有 20 億參數，支援 125 種語言。其訓練分為三階段：自我監督學習、多目標預訓練和微調，提升模型在特定應用場景的性能。Whisper (Radford et al., 2023) 是 OpenAI 開發的多語言 ASR 系統，使用 Transformer 架構，訓練於 68 萬小時音訊數據，擁有 15.5 億參數，支援 50 種語言。本研究選擇了 769 百萬參數的 Whisper Medium，能平衡資源消耗與準確度，適合複雜語音環境。

ASR 模型的糾錯與校正技術：自動語音識別 (ASR) 系統在處理自然語言時，經常會因為口音、背景噪音、語音相似性等問題產生錯誤。為了提高 ASR 系統的準確性，研究者們提出了各種校正技術，包括語音模型優化、語言模型整合以及錯誤修正方法。

Z. Fang 等人提出了非自回歸中文 ASR 錯誤修正方法 PhVEC(Fang et al., 2022)，其創新在於引入音韻字符 (Phonological Tokens)，將拼音作為特殊標記插入原句，提升修正準確性。該方法使用錯誤偵測網絡標記錯誤字符，並生成相應拼音進行修正，如將「你表難過」中的「表」修正為「不要」。實驗顯示，PhVEC 顯著降低字錯率 (WER)，推理速度比

傳統方法快 6.2 倍，大幅提升中文 ASR 錯誤修正的效率和準確性。

Mani et al. (2020) 提出利用機器翻譯模型進行 ASR 錯誤校正和領域適應的方法，透過生成大量錯誤-正確句子對來訓練模型，學習從錯誤輸出到正確文本的轉換規則。實驗顯示，該方法在 Google ASR 系統中顯著提升準確性，字錯誤率 (WER) 改善 7%，BLEU 分數提升 4 分，展現了在領域不匹配情況下的校正效果。

Udagawa et al. (2022) 探討利用大規模語言模型 (如 GPT-2、BERT、RoBERTa) 對 ASR 系統生成的多個候選結果 (N-best hypotheses) 進行重新排序 (rescoring)。模型計算 ASR 的分數 (Score_AM) 和語言模型的分數 (Score_LM)，並以線性組合得到最終得分。實驗結果顯示，雙向語言模型 (如 BERT、RoBERTa) 能顯著改善 ASR 表現並有效減少錯誤率。

Udagawa et al. (2022) 與本篇論文所提出的「BERT 輔助的遮罩語言模型 ASR 校正法」最大的差別在於，(Udagawa et al., 2022) 的方法中，所有的候選結果仍然是基於 ASR 系統的輸出，語言模型 (LM) 僅作為輔助工具，用來計算 ASR 給出的 N-best hypotheses 並對其進行重新排序 (rescoring)。因此，最終的輸出仍然依賴於 ASR 系統曾見過的字彙和詞彙量。相比之下，本研究的校正方法能充分利用語言模型所掌握的詞彙量以及有在醫療情境下經過微調的優勢，直接修正 ASR 系統中因未見過的詞彙或專業術語而產生的錯誤，從而有效減少對專有名詞和專業術語的辨識錯誤。

2.2 Medical Bert 語言模型：MedCKIPBERT

為了實現我們所提出如圖 1 的「利用遮罩增強的語言模型校正技術」(Masking-Enhanced Language Model Correction Technique)，我們需要一個微調 (fine-tune) 在醫療領域上的 BERT，為此我們蒐集醫療文本以作為微調資料。

ChiMed250 醫療文本：我們從網路上各個論文與相關競賽中蒐集了不同的中文醫療資料集，最終彙整成一共 250MB 的醫療文本 (ChiMed250)，其中包含了線上問診對話、醫療詞彙、醫護交班病歷與電子病歷所組成，如圖 2。要特別強調的是，ChiMes250 文本中不管是來自線上資料或是病歷都不包含可識別的個人穩私資訊。

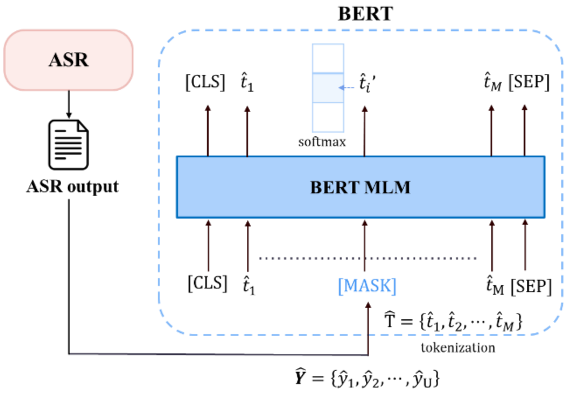


圖 1: 利用遮罩增強的語言模型校正技術

其中有許多語料庫來自於中文醫療資訊處理挑戰榜 CBLUE (Chinese Biomedical Language Understanding Evaluation)。CBLUE 是由中國中文資訊學會醫療健康與生物資訊處理專業委員會在合法開放共享的概念下發起，並由阿里雲天池平台承辦。這個標準旨在推動中文醫學自然語言處理技術的發展，涵蓋命名實體識別、知識抽取、診斷標準化、句子分類以及線上輔助醫療系統的評測 (https://tianchi.aliyun.com/dataset/95414?lang=zh-cn) 。

醫療文本 (ChiMed250) 分成兩大類，(1) 249MB 的門診諮詢文本；(2)1.3MB 的醫護交班病歷：

(1) 門診諮詢文本： 由線上問診對話與醫療詞彙所組成。

- ChiMed (Tian et al., 2019)：醫療保健平台問答，由從線上醫療保健平台“39 健康網”收集的問答對組成，包含 46,731 個問題及 91,416 個回答 (https://www.39health.com.tw/https://www.39health.com.tw/)。
- CHIP-CDN (CBLUE)：病歷中的診斷實體，關於同一種診斷、手術、藥品、檢查、化驗、症狀等往往會有數百種不同的寫法。標準化為臨床上各種不同說法找到對應的標準說法。
- CHIP-CDN_手術詞 (CBLUE)：手術詞彙，關於同一種手術詞彙標準化為臨床上各種不同說法找到對應的標準說法。
- CHIP-CDN_國際疾病臨床詞 (CBLUE)：關於同一種診斷、手術、藥品、檢查、

門診諮詢文本		
資料集	資料類型	資料量
ChiMed	醫療保健平台問答	184,471 KB
CHIP-CDN	病歷中的診斷實體	1,259 KB
CHIP-CDN_手術詞	手術詞彙	116 KB
CHIP-CDN_國際疾病臨床詞	國際疾病臨床詞彙	1,382 KB
CHIP-MDCFNPC	線上問診對話	18,868 KB
IMCS-V2	線上問診對話	16,981 KB
MedDG	問診對話	27,386 KB
THUOCL	醫學詞彙	384 KB
		總資料量: 249.527 MB

醫護交班病歷		
資料集	資料類型	資料量
CHIP-CDEE	中文電子病歷	272 KB
Text2DT	診療指南、診療本文	95 KB
GPT-Mes2023	醫護交班病歷	953 KB
		總資料量: 1.320MB

圖 2：ChiMed250 醫療文本

化驗、症狀等往往會有數百種不同的寫法。標準化國際疾病臨床詞彙為臨床上各種不同說法找到對應的標準說法。

- CHIP-MDCFNPC (CBLUE)：春雨醫師網站 (https://m.chunyuyisheng.com/) 的線上問診資料，訓練集：5,000，驗證集：1,000，測試集：2,000。
- IMCS-V2 (CBLUE)：線上問診對話，包含 10 種兒科疾病，訓練集：2,472，驗證集：833，測試集：811。
- MedDG (CBLUE)：問診對話，醫生和患者交流的句對話歷史，訓練集：17,864，驗證集：2,747，測試集：1,551。
- THUOCL(https://github.com/thunlp/THUOCL)：醫學詞彙，是由中國清華大學自然語言處理與社會人文計算實驗室整理推出的一套中文詞庫，其中也包含了大量醫學類詞彙，醫學詞條數量共 18,749 條。

(2) **醫護交班病歷**：由醫護交班病歷、電子病歷所組成。

- CHIP-CDEE (CBLUE)：中文電子病歷，共 2,485 份電子病歷。
- Text2DT (CBLUE)：臨床診療指南、醫師診療文本，訓練集：300，驗證集：100，測試集：100。
- GPT-Mes2023：將本實驗室與成大醫院的 28 位女性護理師合作蒐集編撰，的醫護交班病歷 M2023 中的訓練集，輸入給 GPT 使其模仿交班病歷特殊形式，生成內容不同的資料。

上述收集的醫療文本中，許多資料集都來自於中國的資源。考量兩岸用詞差異，我們蒐集樂詞網 (https://terms.naer.edu.tw/download/) 上的繁簡醫療詞彙對照表，並對資料進行了轉換。然而，即使經過轉換，仍有部分中國特有的醫療詞彙保留了下來，這些詞彙提供了不同地域性資料和專業術語，進一步豐富了我們的語料庫，並為系統引入了更多新的詞彙和表達方式。

我們將 ChiMed250 語料庫與 BioBERT (Jinhyuk, 2020) 和 EMBERT (Cai et al., 2021) 所使用的語料庫進行比較。BioBERT 的預訓練語料庫規模龐大，超過 21 億字，但全部為英文資料。相比之下，EMBERT 使用 5GB 的中文醫療預訓練資料，主要來自丁香園 (https://www.dxy.cn/) 的醫療問答和論壇數據，專門針對中文醫療場景。

儘管 ChiMed250 語料庫僅有 250MB，其資料來源更專業且多元，涵蓋門診諮詢、線上問診、醫療詞彙、醫護交班記錄和電子病歷等專業醫療文本，覆蓋多個科別及多樣化病歷描述。相比 EMBERT 偏重醫療問答，ChiMed250 提供更廣泛的醫療文本，更全面覆蓋不同的醫療應用場景。

Pretrained BERT CKIP：在獲得醫療文本 (ChiMed250) 後，我們選擇中研院文詞知識庫小組提出的 CKIP BERT (bert-base-chinese) 模型進行微調 (Fine-tuning)。該模型包含 1 億 2 百萬個參數，訓練於中文維基百科和中央通訊社的新聞資料上。表 1 列出了醫療文本資料的大小。

Corpus	Lines	Characters
ChiMed250	1,895,834	87,375,779

表 1：ChiMed250 醫療文本資料統計

MLM 微調 (Fine-tuning)：在具備 250MB 醫療文本、CKIP BERT 預訓練模型和 WordPiece 分詞器後，我們進行微調。即在特定數據集上使用 Masked Language Model (MLM) 技術再訓練 BERT，讓模型更適應特定語言、術語或場景。過程中隨機遮罩部分醫療文本，使用 BERT 進行預測，計算預測結果與原始文本間的損失，並使用梯度下降法更新參數。微調後，模型能更準確處理特定領域語言特徵。圖 3 為 MLM 微調演算法。

Algorithm 1 *BERT Model Fine-tuning with MLM*

Require: *Pretrained BERT model* B_p *, Training dataset* D_{train}
Ensure: *Fine-tuned model* B'
for *each selected* $\mathbf{s}$ *in* D_{train} ***do***
  Generate masked of s_i *, denoted* $M(s_i)$
  Compute predictions $s_i' = M(s_i)$
  Compute loss $L(s_i', s_i)$ *comparing* s_i' *and* s_i
  Update B_p *using* L *to minimize the loss*
end for
$B' \rightarrow B_p$ // *The fine − tuned model*
return B'

圖 3：MLM 微調 (Fine-tuning) 演算法

具體步驟如下：

1. 訓練資料前處理與遮罩：對於一組訓練資料，首先將句子 $S = \{s_1, s_2, \ldots, s_n\}$ 進行前處理。在過程中會隨機選擇其中 15% 的字符 s_i 進行遮罩處理。在這些被選中的字符中，80% 的字符會被替換為 [MASK] 符號，以 $M(s_i)$ 表示，10%會替換為隨機單詞，另外 10%會保持不變。
2. 模型輸入與預測：將遮罩處理後的句子 $S = \{s_1, s_2, M(s_i), \ldots, s_n\}$ 輸入模型，模型根據上下文訊息和已知詞彙預測被遮罩字符的正確內容，計算預測結果s'_i與真實答案 s_i 的交叉熵損失 (Cross Entropy Loss) $L(s'_i, s_i)$。
3. 損失計算與權重更新：計算預測結果與原始文本之間的損失，並使用梯度下降法 (Gradient Descent) 和反向傳播算法 (Backpropagation) 更新模型參數，以最小化損失 L。
4. 迭代訓練：重複上述步驟，直到模型收斂。

透過將 CKIP 的 pretrained BERT 做 MLM 微調在醫學文本上後，就得到了 MedCKIPBERT。

2.3 CER 與 KER 績效指標

在本研究的所有實驗中將會使用兩個不同的指標：CER 與 KER 來對語音辨識模型進行效能評測。

字符錯誤率 (Character Error Rate, CER)：在醫療語音中，講者常混合使用中文和英文，為了更準確地評估 ASR 績效，我們使用字符錯誤率 (Character Error Rate, CER) 替代傳統的詞錯誤率 (Word Error Rate, WER)。CER 計算每個字符的插入、刪除和替換錯誤，以反映雙語系統在處理中英文混合文本時的真實表現。中文字符以單個字計算，英文則按單音節計算。例如，‘glucose’分為‘glu’和‘cose’，與中文的‘血糖’等權重。此外，標點符號也被視為一個字符進行計算，但我們也使用 CER_NP 來表示不計入標點符號的 CER。

關鍵詞錯誤率 (Keyword Error Rate, KER)：在醫療領域，關鍵詞的準確識別比整體翻譯的相似度更為重要，因此我們使用關鍵詞錯誤率 (Keyword Error Rate, KER) 來取代 BLEU 指標作為 ASR 的績效評估標準。KER 類似於詞錯誤率 (Word Error Rate, WER)，但專注於特定的醫療關鍵詞，計算這些關鍵詞的插入、刪除和替換錯誤率。KER 越低，表示辨識結果越好。為確保評估的準確性，我們彙整了一份醫療關鍵詞表，通過初步篩選、使用 ChatGPT 提取醫療詞彙，並由人工確認後，最終確立這些關鍵詞。

3 MLM 的方法

在訓練或預訓練 BERT 時，我們會隨機遮罩部分字符，強迫編碼器根據前後文預測這些字符並調整其編碼，過程包括遮蔽、預測與調整。然而，本論文的「利用遮罩增強的語言模型校正技術」中，MLM 使用專門訓練醫療文本的 Medical BERT，僅進行字符預測，過程為遮蔽與替代。我們進一步區分為常規遮罩和近音遮罩兩種方式。

3.1 常規遮罩 (Masking)

得到 MedCKIPBERT 後我們就可以，使用本研究提出的 BERT 輔助的遮罩語言模型 ASR 校正法，也就是 MLM 校正，此過程不涉及再訓練，而是使用已經訓練好的 BERT 模型來識別並替換文本中可能的錯誤。這個過程將某些

詞替換為[MASK] 標記，然後讓模型預測最可能的替代詞。通過這種方式，可以自動校正 ASR 系統生成的文本中的錯誤，特別是對於關鍵字和專有名詞的校正效果尤為明顯。

***Algorithm* 2** MLM Correction Using BERT

Require: *BERT model B , Input text* $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_T\}$
Ensure: *Corrected text* $\hat{Y}'$

```
Select tokens ŷ_i to be masked based on some criteria
for each selected ŷ_i in Ŷ do
    Generate masked of ŷ_i , denoted M(ŷ_i)
    If confidence (ŷ_i') > α :
        M(ŷ_i) = Compute predictions ŷ_i'
    else:
        M(ŷ_i) = ŷ_i' = ŷ_i
end for
Ŷ' = {ŷ_1', ŷ_2', ⋯, ŷ_T'}
return Ŷ'
```

圖 4：MLM 校正 (Correction) 演算法

圖 4 具體步驟如下：

1. 訓練資料前處理與逐步遮罩：將 ASR 輸出的文本 $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_T\}$ 輸入輸入到 BERT 模型中。從序列中的第一個字符 $\hat{y}_1$ 開始做遮罩 $M(\hat{y}_1)$，其餘字符則不動。
2. 預測遮罩：使用 BERT 模型對遮罩後的文本 $\hat{Y} = \{M(\hat{y}_1), \hat{y}_2, \cdots, \hat{y}_T\}$ 進行預測，取最高機率的字符當作預測字符 $\hat{y}_i'$，當預測信心值 $confidence(\hat{y}_i')$ 有大於 α，才將預測結果 $\hat{y}_i'$ 取代 $M(\hat{y}_i)$，否則就保留原來的 token $\hat{y}_i$ 不動。α 在本研究的實驗中設為 0.9。
3. 迭代替代：將預測結果中的替代詞替換回原文本中 $\hat{Y} = \{\hat{y}_1', \hat{y}_2, \cdots, \hat{y}_T\}$，重複上述步驟，直到序列中的每個字符都被遮罩過。

3.2 近音遮罩 (Partial homophones masking)

我們發現在 ASR 的輸出中，多數錯誤源於「同音/近音異字」的錯誤辨識。為了改善這一問題，我們提出了一種近音遮罩技術 (Partial homophones masking)，如圖 5，該技術在遮罩替代時，會優先選取同音或近音中機率最高的作為預測結果。

在預測過程中，系統先判斷字符 (token) 是中文還是英文，並採取不同方法限縮預測範圍至同音/近音異字。對於中文，我們使用 python 函式庫 lazy_pinyin 將中文轉為拼音，並篩選出「同音不同調」的候選字，如「活」(huó) 和「或」、「霍」、「火」等字。對於英文，則使用 Metaphone 函式庫來尋找發音相似但拼法不同的字符，從而提高 ASR 系統的準確性。

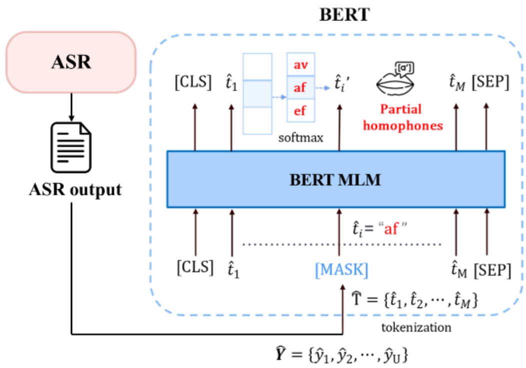


圖 5：近音遮罩框架

4 實驗與結果

為了驗證本論文所提的利用遮罩增強的語言模型校正技術，我們利用 GPT 生成一般醫療情境腳本，請人錄音而集成為醫療語音的測試集，然後利用 Google STT 以及 Whisper Medium 作 ASR 測試，然後對照使用 MedCKIPBERT 的語言模型進行常規遮罩以及近音遮罩的校正效果。

4.1 測試集

模擬醫療語音場景設定在包含專業術語的醫療對話中，由 3 位錄音員錄製了 354 句音檔，這些句子從 ChiMed250 的門診諮詢文本中選取，並利用 GPT 生成風格相似但內容不同的文本，如表 2 所示。有關利用 ChatGPT 生成技術，能利用有限範例文本生產具多樣性同風格的擬真病歷可參考 (Chung et al., 2024)。如此生成的內容保留醫療關鍵字，並確保測試集的獨立性和結果的可靠性，當然也不涉及任到病患的穩私性資訊。

一般醫療文本	鉤端螺旋體病會導致腸道黏膜的損傷，引起蛋白的吸收減少，另外還會導致貧血，會導致低蛋白水腫的情況，所以需要給予抗寄生蟲治療的同時還需要給予支援性的治療，比如給予大量的優質蛋白的攝入，另外需要補充大量的維生素和鐵既有利於貧血的改善，病情就可以逐漸的恢復。
GPT 生成風格相似文本	鉤端螺旋體病引致的腸道黏膜受損，會嚴重影響蛋白質的正常吸收過程，並可能導致貧血以及低蛋白血症引起的水腫現象，在針對寄生蟲的抗感染治療外，強烈建議配合相應的支持性治療措施。

表 2：使用 GPT 生成風格相似的文本

4.2 實驗結果

我們比較了 Google 和 Whisper Medium 的 ASR 系統在醫療語境中的辨識績效，並使用 MedCKIPBERT 進行遮罩增強語言模型校正。MedCKIPBERT 是在 CKIP BERT 基礎上經 WordPiece 切詞微調而成。我們測試了兩種校正技術：(1) 常規遮罩：取最高機率的字符作為預測，以提升準確性；(2) 近音遮罩：根據發音從同音不同調或類似發音候選中取最高機率字符。分別以 mask-MedCKIPBERT 和 homo-MedCKIPBERT 代表。

一般醫療語音	
Google	Whisper Medium
CER: 12.46% KER: 15.37%	CER: 9.62% KER: 20.82%
CER_NP: 8.36% CER_NP -CER: 4.10%	CER_NP: 6.06% CER_NP -CER: 3.56%
mask-CKIPBERT	
CER: 12.42% (-0.04) KER: 15.09% (-0.28)	CER: 9.92% (+0.30) KER: 20.82% (-0.00)
CER_NP: 8.28% (-0.08) CER_NP -CER: -4.14%	CER_NP: 6.22% (-0.16) CER_NP -CER: -3.70%
homo- CKIPBERT	
CER: 12.41% (-0.05) KER: 15.09% (-0.28)	CER: 9.91% (+0.29) KER: 20.82% (-0.00)
CER_NP: 8.26% (-0.10) CER_NP -CER: -4.15%	CER_NP: 6.21% (-0.15) CER_NP -CER: -3.70%

表 3：使用 CKIPBERT 串接 ASR 在一般醫療語音的組態效果總覽

首先比較了兩種不同的自動語音識別系統：Google、Whisper Medium 的測試結果，顯示如上表 3 的第一、二列。Google 儘管 CER 較差，但在關鍵詞錯誤率 (KER) 為最佳。相較而言，雖然 Whisper Medium 在 CER 上效果較佳，但其 KER 較差則顯示出在醫療關鍵詞的識別上明顯較弱。接下來加入語言模型的校正機制，先使用中研院做的 CKIPBERT 不經過微調，直接串接「BERT 輔助的遮罩語言模型 ASR 校正法」進行常規遮罩 (mask) 與近音遮罩法 (homo) 校正，可以看到效果有些微提升、甚或有些微降低，校正效果並不顯著。

CER_NP 是去除標點符號後的 CER，旨在更精確評估 ASR 系統處理純文字的表現。“CER_NP - CER”代表去除標點前後的 CER 差異。我們觀察到 Google 和 Whisper Medium ASR 常無法準確識別標點符號，導致後續 BERT 修正困難。因此，我們分析了去除標點後的字符錯誤率 (CER_NP)。

相較而言，表 4 為使用本研究在醫學領域下建構的 BERT 模型 MedCKIPBERT 來做校正，透過各種組態校正後，Google、Whisper Medium 各的 CER 和 KER 相較於表 3，只用一般文本的語言模型訓練的 CKIPBERT 來作校正的效果，均能有所改善。

一般醫療語音	
Google	Whisper Medium
CER: 12.46% KER: 15.37%	CER: 9.62% KER: 20.82%
CER_NP: 8.36% CER_NP -CER: 4.10%	CER_NP: 6.06% CER_NP -CER: 3.56%
mask-MedCKIPBERT	
CER: 12.00% (-0.46) KER: 13.64% (-1.73)	CER: 9.10% (-0.52) KER: 16.74% (-4.08)
CER_NP: 7.89% (-0.47) CER_NP -CER: -4.11%	CER_NP: 5.58% (-0.48) CER_NP -CER: -3.52%
homo- MedCKIPBERT	
CER: 11.97% (-0.49) KER: 13.64% (-1.73)	CER: 8.96% (-0.66) KER: 15.80% (-5.02)
CER_NP: 7.81% (-0.55) CER_NP -CER: -4.16%	CER_NP: 5.29% (-0.77) CER_NP -CER: -3.67%

表 4：使用 MedCKIPBERT 串接 ASR 在一般醫療語音組態總覽

在 Whisper Medium 經 homo-MedCKIPBERT 校正後，CER 與 KER 的表現效果最佳，CER 表現下降了 0.66%，CER_NP 表現下降了 0.77%， KER 下降了 5.02%。原本辨識效果就不錯的 Google 經過校正後也均能有更好的改善，在 homo-MedCKIPBERT 中效果為最佳：CER 下降了 0.49%，CER_NP 下降了 0.55%，KER 下降了 1.73%，但與一般遮罩的 mask-MedCKIPBERT 比起來，在統計上並沒有顯著的差異。相同的觀察也發生在表 3。

綜合來看，儘管我們的猜測大多數 ASR 錯誤來自於同音/近音異字，但近音遮罩與一般遮罩在校正 ASR 輸出效果上的差異，並不顯著。反而是 MedCKIPBERT 使用了大量預訓練文本切割的 WordPiece 提供更大的字典，這種分詞方法能捕捉更多語義信息，使得校正的效果優於僅用一般文本訓練的 CKIPBERT 模型的校正效果，有助於更準確地識別和校正字符錯誤，特別是與醫療術語相關的 KER 績效上能表現更好。

Ground Truth	盆腔炎是指女性盆腔內的生殖器官和周圍結締組織，包括子宮、輸卵管、卵巢以及盆腔腹膜等，發生炎症的情況。
Google CER: 12.46% KER: 15.37%	盆腔炎是指女性盆腔內的生殖器官和周遭結締組織，包括子宮，輸卵管，卵巢以及盆腔腹膜的，發生癌症的情況。
Google: homo- MedCKIPBERT CER: 11.97% (-0.49) KER: 13.64% (-1.73) CER_NP: 7.81% (-0.55) CER_NP -CER: -4.16%	盆腔炎是指女性盆腔內的生殖器官和周遭結締組織，包括子宮，輸卵管，卵巢以及盆腔腹膜的，發生炎症的情況。

表 5：使用 homo-MedCKIPBERT 串接 Google 在一般醫療語音實例

Ground Truth	盆腔炎是指女性盆腔內的生殖器官和周圍結締組織，包括子宮、輸卵管、卵巢以及盆腔腹膜等，發生炎症的情況。
Whisper Medium CER：9.62% KER：20.82%	盆腔炎是指女性盆腔內的生殖器官和周遭結地組織，包括子宮，舒卵管，卵巢以及盆腔附膜等，發生炎症的情況，
Whisper Medium: homo- MedCKIPBERT CER: 8.96% (-0.66) KER: 15.80% (-5.02) CER_NP: 5.29% (-0.77) CER_NP -CER: -3.67%	盆腔炎是指女性盆腔內的生殖器官和周遭結締組織，包括子宮，輸卵管，卵巢以及盆腔腹膜等，發生炎症的情況，

表 6：使用 homo-MedCKIPBERT 串接 Whisper Medium 在一般醫療語音實例

4.3 實例

以實際辨識測試中發生的句子為例，我們用三種不同顏色分別表達對同類型的錯誤：替換：紅色、插入：綠色<>、刪除：藍色刪除線、Keyword：螢光。在還沒做任何校正前，Google 與 Whisper Medium 在 CER 上的表現相差不多，然而在 KER 方面，Google 的表現明顯更佳。表 5 與表 6，分別為 Google 與 Whisper Medium 使用 homo-MedCKIPBERT 後的校正結果。CER 分別降低了 0.49% 與 0.66%，而 KER 的改善則更顯著，分別降低了 1.73% 與 5.02%。

5 結論

我們在實驗過程中觀察到：首先，在現有 AI 大廠於醫療語音識別績效的比較上，不論是一般語音還是病房交班語音，Whisper Medium 在字元錯誤率 (CER) 方面表現較佳，而 Google 在關鍵詞錯誤率 (KER) 方面更優。這反映了 Whisper Medium 擁有更好的聲學模型 (AM，Acoustic model) 和發音模型 (PM，Pronunciation model)，而 Google 則具備更強的語言模型 (LM，Language model)，這可能是由於其接觸了更多的醫學文本。

其次，MLM 的校正機制在都能改進一般 ASR 的績效，特別是在特殊專業領域中關鍵詞錯誤率 (KER) 方面。其根本原因在於，專業領域語音辨識的最大挑戰在於隱晦的術語，這些術語很難全面地收錄於訓練語音集中，這不僅增加了識別的難度，還會影響前後文的識別判斷。本論文提出的 MLM 校正機制有效提升了關鍵詞的辨識效果。相比於語音庫取得成本高，較全面性覆蓋專業領域術語的文本取得成本低廉，且語言模型的非監督式訓練成本也較低。

總結來說，本研究針對 ASR 在專業場域中的辨識績效欠佳以及專業領域語料庫蒐集困型校正技術」。該方法透過在 ASR 後串接針對特殊應用場域的語言模型作為遮罩語言模型進行校正，在特殊專業領域中有效降低了字

符錯誤率 (CER) 與關鍵詞錯誤率 (KER)，顯著減少了專有名詞和專業術語的辨識錯誤。

本研究的測試語音主是來自線上醫療文本，再經過生成式 AI 的 ChatGPT 所產生的腳本，再由人錄音而成。未來研究方向是針對門診中，醫師對病患的問認與應答，以及病房中之護理交班或是醫師巡房問答等醫療情境中的語音進行測試。

References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., . . . Chen, G. (2016). *Deep speech 2: End-to-end speech recognition in english and mandarin.* Paper presented at the International conference on machine learning.

Cai, Z., Zhang, T., Wang, C., & He, X. (2021). *EMBERT: A pre-trained language model for Chinese medical text mining.* Paper presented at the Web and Big Data: 5th International Joint Conference, APWeb-WAIM 2021, Guangzhou, China, August 23–25, 2021, Proceedings, Part I 5.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). *Listen, attend and spell: A neural network for large vocabulary conversational speech recognition.* Paper presented at the 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP).

Chung, S.-L., Fan, J.-H., & Ting, H.-W. (2021). *Chinese Medical Speech Recognition with Punctuated Hypothesis.* Paper presented at the Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021).

Chung, S.-L., Lin, J.-Y., & Li, X.-Q. (2024). *Applying Generative Language Models to Generate Synthetic Medical Records: ChaVinci.* Paper presented at the Paper presented at the Proceedings of the 36th Conference on Computational Linguistics and Speech Processing (ROCLING 2024).

Enarvi, S., Amoia, M., Teba, M. D.-A., Delaney, B., Diehl, F., Hahn, S., . . . Pinto, J. (2020). *Generating medical reports from patient-doctor conversations using sequence-to-sequence models.* Paper presented at the Proceedings of the first workshop on natural language processing for medical conversations.

Fang, Z., Zhang, R., He, Z., Wu, H., & Cao, Y. (2022). *Non-Autoregressive Chinese ASR Error Correction with Phonological Training.* Paper presented at the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Jinhyuk, L. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*, 1234.

Mani, A., Palaskar, S., Meripo, N. V., Konam, S., & Metze, F. (2020). *Asr error correction and domain adaptation using machine translation.* Paper presented at the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). *Robust speech recognition via large-scale weak supervision.* Paper presented at the International Conference on Machine Learning.

Tian, Y., Ma, W., Xia, F., & Song, Y. (2019). *ChiMed: A Chinese medical corpus for question answering.* Paper presented at the Proceedings of the 18th BioNLP Workshop and Shared Task.

Udagawa, T., Suzuki, M., Kurata, G., Itoh, N., & Saon, G. (2022). Effect and analysis of large-scale language model rescoring on competitive asr systems. *arXiv preprint arXiv:2204.00212*.

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*.

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., . . . Wang, G. (2023). Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

A Chinese Education Broadcast Emotion Corpus
(中文教育廣播情緒語料庫)

Pin-Hsiu Lin, Hou-Chiang Tseng, Kuan-Yu Chen

National Taiwan University of Science and Technology
{torrance_lin, tsenghc, kychen}@mail.ntust.edu.tw

摘要

本研究提出了一個中文教育廣播之情緒語料庫，收錄台北酷課雲裡小學一年級至高中三年級各科目的教學影片、動畫片、幼兒學習影片、趣味科學節目等，共計 3,060 部影片。我們採用 Azure 語音辨識服務，產生影片中語音內容的轉寫文字；使用 Hume 語音情緒辨識服務，為影片內每個句子產生對應的情緒標籤。在實驗中，我們在中文教育廣播之情緒語料庫上進行一系列的實驗，並探討 8 組常用的語音特徵在語音情緒辨識的任務成效。這些實驗成果，為中文教育廣播之情緒語料庫奠定了重要的基礎，並可作為後續相關研究的參考。

Abstract

This study presents a Chinese Education Broadcast Emotion Corpus (CEBE), which includes a total of 3,060 videos from the Taipei City CooC Cloud platform, ranging from first grade in elementary school to twelfth grade in high school. The videos encompass teaching videos, animated clips, early childhood learning videos, and science programs. We utilized the Azure Speech Recognition service to generate transcriptions of the speech content in the videos and employed the Hume Speech Emotion Recognition service to assign corresponding emotion labels to each sentence in the videos. In our experiments, we conducted a series of tests on the CEBE corpus, exploring the effectiveness of eight commonly used speech features in speech emotion recognition. These experimental results establish an important foundation for the CEBE corpus and serve as a reference for future related research.

關鍵字：中文教育廣播、情緒語料庫、語音情緒辨識

Keywords: Chinese Education Broadcast, Speech Emotion Corpus, Speech Emotion Recognition

1 緒論

情緒辨識技術是人工智慧中人機互動領域裡一個重要的研究方向。傳統上，情緒是人類交流中不可或缺的一部分，我們會透過語言、表情和肢體動作表達和感知彼此的情緒。而隨著科技的日新月異，人工智慧不僅能夠理解和回應人類語言，還能夠識別和解讀人類情緒。它能透過分析臉部表情、語音語調、生理訊號等多種方式，將抽象的情緒轉化為可量化的數據。這一技術的發展帶來了眾多潛在應用，從智能客服和個人助理到心理健康監測和市場分析，都能看到情緒辨識的身影。

聲音中所蘊含的音調、音量、節奏等各種特徵，是情緒辨識中重要的資訊之一，可以用來識別說話者的情緒狀態（R. A. Calvo et al., 2010）。在對話的交流中，通常被區分為 Primary Channel 和 Secondary Channel（S. Casale et al., 2008）。前者與語法（Syntactic）、語意（Semantic）相關，傳達語言訊息；後者傳達副語言訊息，如語調（Tone）、情緒狀態（Emotional State）及手勢等。通過副語言訊息，可以更加的理解說話者當下的狀態，避免有誤解、判斷錯誤的情況發生。

隨著深度學習的演進，語音情緒辨識領域取得了顯著進展。然而，語音情緒辨識的訓練仍面臨諸多挑戰。除了資料集難以收集和需要大量人力時間進行標註外，不同文化、背景和個體之間的語音情緒表達方式存在差異。此外，如何保證語音情緒辨識系統的實時性和可靠性也是關鍵挑戰。

本研究建立了一個中文教育廣播情緒語料庫，涵蓋台北酷課雲的小學至高中教學影片、動畫、幼兒學習影片等，共收錄 3,060 部影片。我們使用 Azure 語音辨識服務轉寫影片語音內容（Microsoft, n.d.），並透過 Hume 語音情緒辨識服務為每個句子標記情緒（Hume AI, n.d.）。此外，我們更探討了 8 組常見的語音特徵在語音情緒辨識中的效果。這些實驗成果，為中文教育廣播之情緒語料庫奠定了重要的研究基礎，並可作為後續相關研究的參考。

2 文獻

在語音情緒辨識（Speech Emotion Recognition, SER）領域，過去的研究主要集中於情緒分類的問題。這些研究通常採用各種機器學習和深度學習方法，利用聲學特徵來對語音樣本進行情緒分類。

早期的研究多使用 LSTM 或 CNN 作為深度學習的架構，後期也出現了兩者一起使用的模型。如 Jianfeng Zhao 等人（2019）構建了兩個網路模型：一個是一維 CNN-LSTM，用於從原始音頻片段中學習特徵；另一個是二維 CNN-LSTM，用於從對數梅爾頻譜（Log-mel Spectrogram）中學習特徵。該實驗使用 Berlin Emotional Database（Burkhardt et al., 2005）和 IEMOCAP（Busso, C. et al., 2008）這兩種資料集做驗證。在 Speaker-dependent 的條件下，一維 CNN-LSTM 和二維 CNN-LSTM 分別在 Berlin Emotional Database 達到了 92.34% 和 95.33% 的平均準確率；而在 IEMOCAP 上一維 CNN-LSTM 和二維 CNN- LSTM 則分別達到了 67.92% 和 89.16% 的平均準確率。研究結果表明，結合 CNN 和 LSTM 的深度學習網路能夠從語音數據中學習到豐富的情緒特徵，並顯著提高語音情緒辨識的準確性。

隨著注意力機制（Attention Mechanism）的崛起，有效的解決了 LSTM 或 CNN 因為應付較長的序列導致微分時梯度消失、梯度爆炸的問題，也因為可平行運算的模型設計，大大的減緩計算時間隨序列長度成正比增加的問題。Y. Wang 等人（2021）使用了在語音辨識（Automatic Speech Recognition, ASR）領域中性能最優秀的自監督式預訓練模型 Wav2Vec 2.0 和 HuBERT，探討模型的部分微調（僅調整 Transformer 層）及全數微調（同時調整 CNN 和 Transformer 層），在 SER、Speaker Verification（SV）和 Spoken Language Understanding（SLU）任務的成效。在語音情緒辨識的研究中，該論文以 IEMOCAP 資料集進行實驗。結果顯示微調模型顯著地優於未微調模型；而部分微調的 HuBERT 大型模型表現最佳，表明這種方法在語音情緒辨識任務中特別有效。

相較於較制式的情緒種類，Russell, James 和 Mehrabian, Albert（1977）提出了 AVD 模型，它是一種捕捉人類情緒體驗豐富性的強大工具。通過將情緒分解為這三種維度：Arousal（激活度，從平靜到興奮，描述情緒的強度或激烈程度）、Valence（愉悅度，從不愉快到愉快，描述情緒的正面或負面程度）及 Dominance（支配度，從被動到主動，描述感受到的控制或影響力程度）。研究人員和實踐者可以更深入地了解情緒的性質和動態，從而促進更好的情緒理解和調節。

此外，傳統的方法依賴於靜態描述（例如統計函數或通用背景模型），但這些方法無法有效地捕捉語音表達中的動態時間變化。近期的深度學習方法可以直接提取句子級別中對應的特徵，但由於每個語音語句的長度不同，因此常見的方式會先將語句進行裁剪，或在最後填補零的方式，將所有語句的長度對齊，但這些方式都可能會損失有用的資訊。有鑒於此，W. -C. Lin and C. Busso（2023）提出了一種新的動態區塊分割法，以事先定義區塊個數的方式，動態地將每個語音語句切割成固定個數的區塊。爾後，將每個區塊分別進行特徵的擷取，最後再進行整合。透過這樣的技術，期望可以完整地、不破壞語音內容完整性地進行語音資訊的特徵擷取，進而提升情緒辨識的效能。

3 中文教育廣播之情緒語料庫

我們提出了一個中文教育廣播情緒語料庫（A Chinese Education Broadcast Emotion Corpus, CEBE），這是針對教育領域專門設計的語料庫，旨在促進情緒辨識與教育科技的結合。我們統合了台北酷課雲小學一年級至高中三年級各科目教學影片、動畫片、幼兒學習影片、趣味科學節目等，共計 3,060 部的影片，涵蓋了廣泛的教育資源和年齡層。透過分析教育影片中的情緒表達，研究者可以探索教

Tag	Amount	Tag	Amount	Tag	Amount
Interest (O)	122,500	Sadness (S)	1,354	Negative Surprise (U)	97
Concentration (O)	73,093	Sympathy (O)	955	Horror (F)	90
Calmness (N)	38,640	Disgust (D)	561	Love (O)	77
Determination (O)	22,210	Positive Surprise (U)	534	Adoration (O)	44
Excitement (H)	13,830	Awkwardness (O)	526	Embarrassment (O)	33
Contemplation (O)	8,776	Satisfaction (H)	522	Relief (O)	20
Joy (H)	8,554	Tiredness (O)	470	Awe (O)	17
Amusement (H)	5,557	Nostalgia (O)	400	Desire (O)	11
Realization (O)	5,231	Pride (O)	315	Triumph (H)	4
Admiration (O)	5,163	Anxiety (F)	280	Guilt (O)	3
Anger (A)	4,106	Craving (O)	267	Shame (O)	3
Confusion (O)	3,710	Pain (S)	215	Envy (O)	2
Doubt (O)	3,634	Empathic Pain (S)	199	Ecstasy (H)	1
Distress (S)	2,966	Boredom (O)	180	Contentment (H)	0
Aesthetic Appreciation (O)	1,870	Contempt (C)	168	Entrancement (O)	0
Disappointment (S)	1,511	Fear (F)	145	Romance (O)	0

表 1. CEBE 資料集情緒種類和數量表，按照數量由大到小排序。由於類別數量繁多，因此我們根據語意相似性，重新將情緒類別歸納為 Angry（A）、Sad（S）、Happy（H）、Surprise（U）、Fear（F）、Disgust（D）、Contempt（C）、Neutral（N）、Other（O）與 No agreement（X），共十種。

師的情緒如何影響學生的學習效果，以及學生在不同情境下的情緒反應，抑或是學生在進行翻轉教育時，在台上的表達能力是否良好。此外，CEBE 可用於訓練情緒辨識模型，這些模型可以應用於各種教育科技中，例如智能教室、虛擬教練、情緒回饋系統等。

在資料標註方面，我們首先使用 Azure 取得這些影片的逐字稿，精確地將影片中的語音內容轉換為文本。接著，我們使用 Hume Python SDK 對每部影片中每段句子的音檔進行情緒分析，產生 48 種情緒維度。這些情緒維度代表了人們根據語音變化能夠區分的情緒含義，它是根據模型對韻律的分析來判斷的機率。換句話說，當某個音檔的「開心」維度得分最高時，這表明該音檔最有可能被人們解讀為表達開心的情緒。因此，我們選擇最高分的情緒維度作為該句子的情感標記結果。最後將文本、分數及最終結果合併成一個資料表。

為證明 Hume Python SDK 標註的可靠性，Alan S. Cowen 和 Dacher Keltner（2021）的研究使用大量自然資料和機器學習技術來自動進行情感標記，成功識別多達 34 種情感。透過 fMRI 驗證，發現這些自動標記的情感與大腦的神經活動模式高度一致，且能跨文化、跨模式準確預測情感反應。結果顯示，機器學習自動標記能捕捉到比傳統模型更複雜的情感表徵，證明其在情感研究中的有效性。

此外，我們進行了嚴格的資料清理以確保語料庫的品質和一致性，包括移除重複的影片、長度低於兩秒的語句、與資料庫不符或不存在的影片等。最終，我們將語料庫切分為 328,844 個句子，總長約 372 小時又 10 分鐘。CEBE 資料集所包含的情緒種類和數量如表 1 所示。由於中文教育廣播情緒語料庫涵蓋了豐富的情感表達形式，並提供大量、充足的資料量，將能促進語音情緒辨識模型的研究與發展。

以 emotionResult_100_0_0_0.wav 為例，該音檔是來自編號 100 部影片第 0 大段第 0 小段中的第一句，語者說的話為”台北市線上教學影片”。該音檔經過 Hume Python SDK 得到的 48 種情緒維度中由最高的前三名依次為 Calmness, Boredom 和 Concentration，分數分別為 0.5396, 0.3309 及 0.3285。我們選擇了得分最高了 Calmness 作為此音檔的標註結果。

Layer	Channels	Kernel	Stride	Dimension	Activation
Input	N/A	N/A	N/A	$m \times d$	N/A
Permute	N/A	N/A	N/A	$d \times m$	N/A
CNN-block	128	(1,3)	1	depends	ReLU
CNN-block	64	(1,3)	1	depends	ReLU
CNN-block	32	(1,3)	1	depends	ReLU
Flatten	N/A	N/A	N/A	depends	N/A
Linear	N/A	N/A	N/A	$1 \times b$	ReLU

表 2. 動態區塊分割模塊的模型架構與參數。

4 實驗設定

4.1 聲學特徵

openSMILE 工具（F. Eyben et al., 2010）是一套經常被使用於聲學特徵提取的工具，並且他提供許多公用的語音特徵集。在本研究中，我們使用 openSMILE 工具進行語音特徵的抽取，並採用 INTERSPEECH 國際會議在 2009~2013 年，每年所舉辦的各項語音內涵比賽之特徵集 [1]（Björn Schuller et al., 2009, Schuller et. al., 2010, Björn Schuller et al., 2011, Björn Schuller et al., 2012, B. Schuller et al., 2013），以及常見的語音情緒辨識特徵集 emobase、emoLarge 和 emobase2010 等共八種特徵集，為每個語音語句進行語音特徵的抽取。

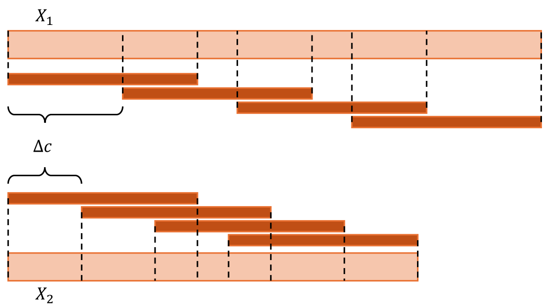


圖 1. X_1與X_2為兩個不同長度的語音語句。當我們將區塊個數參數C設定為5，且固定區塊長度的參數w後，語音語句X_1與X_2都被切割成C個相同長度的區塊。切割時，因為X_1較長，所以X_1的移動時間長度Δc較X_2的長。

4.2 語音情緒辨識模型

由於先前的研究指出，動態區塊分割法可以有效地解決語音語句長度不一的問題，且在語音情緒辨識的任務中能夠取得良好的任務成效（W. -C. Lin and C. Busso, 2023）。因此，本研究以此為基礎，建立一套簡單而有效的語音情緒辨識模型。我們的語音情緒辨識模型有四個模塊：特徵提取、動態區塊分割、區塊層級的特徵表示、句子層級的時間聚合。

特徵提取模塊是用來從語音信號中提取幀（Frame）級別的聲學特徵，如頻譜圖特徵（Spectrogram）等的低階特徵。因此，當給定一個語音訊號X後，特徵提取模塊可以將其表示成一系列的低階特徵向量$F = \{f_1, \dots, f_T\}$。

接著，在動態區塊分割模塊裡，則是要將不定長度的低階特徵向量，轉換成固定個數的高階特徵向量。我們首先定義一個區塊個數參數C，與區塊長度的參數w。每個語音語句不論長度，都會被切割成C個長度w為區塊，且切割時的移動時間長度是Δc：

$$\Delta c = \frac{T-w}{C-1} \quad (1)$$

由式子(1)可知，當區塊長度w固定時，若區塊數量C增加，移動時間長度Δc就會縮小，導致區塊之間的重疊增加。切割的示意圖如圖 1 所示。在切過後，我們使用三層的一維卷積神經網路（1D CNN）來處理每一個區塊內的低階特徵向量。最後，為每一個區塊產生一個高階的特徵向量。詳細的模型參數如表 2 所示。因此，通過動態區塊分割模塊，語音訊號X的低階特徵向量$\{f_1, \dots, f_T\}$，將轉換為C個高階特徵向量$H = \{h_1, \dots, h_C\} \in \mathbb{R}^{C \times b}$。這種動態區塊分割方法特別適合處理長度不定的時

[1] https://audeering.github.io/opensmile/get-started.html#default-feature-sets

間序列資料。由於切割過後片段的大小是固定的，這不僅簡化了深度學習模型的設計，也使得模型在處理不同長度的語句時更加一致；其次，這一過程可以並行化處理，從而允許同時計算多個區塊的特徵表示，大幅提升計算效率。

最後，句子層級的時間聚合模塊則是用來總結區塊層級的特徵，有效地統合整個語句的情緒訊息，從而產生最佳的語音情緒辨識結果。我們探討了平均向量法、權重式平均向量法、注意力法與自注意力法等四種模型架構。

- 平均向量法：將所有區塊的高階特徵向量進行平均，得到最終的語句級別表示向量z：

$$z = \frac{1}{C}\sum_{t=1}^{C} h_t \quad (2)$$

- 權重式平均向量法：在平均向量法中，每個區塊的權重皆是一致的。然而，在一段語句中，情緒的展現可能僅出現在某些片段中而已。因此，在整合區塊的向量表示法時，每個區塊應該有不同的權重。我們首先利用一個簡單的前饋神經網路（Feedforward Neural Network）搭配激活函數Sigmoid，為每個區塊進行權重的計算：

$$g_t = \sigma(W^a \cdot h_t + b) \quad (3)$$

其中，W^a為前饋神經網路的參數，$\sigma(\cdot)$表示Sigmoid 函數，b則為偏移量（Bias）。接著，我們將每個區塊的向量表示法乘上對應的權重後相加，作為最後句子層次的向量表示法：

$$z = \sum_{t=1}^{C} g_t h_t \quad (4)$$

- 注意力法：雖然權重式平均向量法已經考慮了每個區塊應有不同的權重，但權重的計算方式卻是每個區塊獨自運算，而沒有考慮全域（也就是所有區塊）的資訊。有鑒於此，參考先前的研究（T. Luong et al., 2015），我們使用注意力法，藉由考慮全域的資訊，為每個區塊計算權重。首先，我們將C個高階特徵向量輸入一層遞迴式神經網路（Recurrent Neural Network, RNN），為每個高階特徵向量轉換為具有相鄰資訊的向量表示法：

$$H^{RNN} = RNN(H) \quad (5)$$

然後，我們利用具有相鄰資訊的向量表示法來為每個區塊計算注意力權重：

$$s_t = (h_t^{RNN})^\top W^b h_C^{RNN} \quad (6)$$

$$[\alpha_1, \ldots, \alpha_t, \ldots \alpha_C] = softmax([s_1, \ldots, s_t, \ldots s_C]) \quad (7)$$

其中，W^b為一個可訓練的模型參數。藉由注意力權重α_t加權對應的相鄰資訊向量，產生一個統合資訊向量v：

$$v = \sum_{t=1}^{C} \alpha_t\, h_t^{RNN} \quad (8)$$

最後，我們串接統合資訊向量v與最後一個區塊的相鄰資訊向量h_C^{RNN}，通過一個前饋神經網路與激活函數$tanh(\cdot)$來生成最終的句子層級特徵表示：

$$z = tanh\,(W^e[v; h_C^{RNN}]) \quad (9)$$

W^e是前饋神經網路的模型參數。

- 自注意力法是基於現在流行的多頭自注意力（Multi-head Self-attention）結構的模型。他不僅結合了權重式平均向量法與注意力法考慮到為每個區塊賦予不同權重的優點，更進一步地改善在注意力法中，產生相鄰資訊的向量表示法時，僅考慮單向資訊的缺點。此外，自注意力還有可平行化的優點，可以加速整個計算過程。再者，由於多頭的模型設計，可以讓模型自動地從不同面向加權不同區塊的權重。在實際的運算上，我們將C個高階特徵向量送入多頭自注意力模型進行運算：

$$Head_j = selfatt(HW_j^Q, HW_j^K, HW_j^V) \quad (10)$$

其中，$Head_j$表示第j組自注意力運算結果，$selfatt(\cdot,\cdot,\cdot)$是自注意力運算函式，而$W_j^Q$、$W_j^K$和$W_j^V$是自注意力運算中的參數：

MRR	Feature Set							
	IS2009	IS2010	IS2011	IS2012	IS2013	emobase	emoLarge	emobase2010
平均向量法	0.8338	**0.8555**	0.8376	0.8437	0.8526	0.8293	0.8307	0.8267
權重式平均向量法	0.8312	0.8312	0.8378	**0.8478**	0.8321	0.8332	0.8289	0.8468
注意力法	0.8467	0.8325	0.7995	0.8433	0.8451	0.8096	0.8209	**0.8523**
自注意力法	0.8176	0.8375	0.7878	0.8337	0.8013	0.7981	0.7981	**0.8401**

表 3. CEBE 資料集在使用平均倒數排名評估下的實驗結果。

Macro F1-Score	Feature Set							
	IS2009	IS2010	IS2011	IS2012	IS2013	emobase	emoLarge	emobase2010
平均向量法	0.1449	0.1456	0.1473	0.1491	0.1480	0.1327	0.1443	0.0958
權重式平均向量法	0.1455	0.1433	0.1476	0.1491	0.1466	0.1403	0.1451	0.0958
注意力法	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958
自注意力法	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958

表 4. CEBE 資料集在使用 Macro Average F1-score 評估下的實驗結果。

Weighted F1-Score	Feature Set							
	IS2009	IS2010	IS2011	IS2012	IS2013	emobase	emoLarge	emobase2010
平均向量法	0.6970	0.6980	0.6994	0.7010	0.7000	0.6847	0.6968	0.6538
權重式平均向量法	0.6972	0.6960	0.6994	0.7007	0.6992	0.6925	0.6962	0.6538
注意力法	0.6538	0.6538	0.6538	0.6538	0.6538	0.6538	0.6538	0.6538
自注意力法	0.6538	0.6538	0.6538	0.6538	0.6538	0.6538	0.6538	0.6538

表 5. CEBE 資料集在使用 Weighted Average F1-score 評估下的實驗結果。

$$selfatt(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (11)$$

d表示向量的維度。我們將J組自注意力運算結果併合後，透過一個簡單的轉換，形成最終的向量表示法$\widetilde{H} \in \mathbb{R}^{C \times b}$：

$$\widetilde{H} = Concat(Head_1, \dots, Head_J)W^O \quad (12)$$

W^O為簡單轉換的參數矩陣。最後，我們將每個向量表示法相加後取平均，作為最後句子層次的向量表示法：

$$z = \frac{1}{C}\sum_{t=1}^{C}\widetilde{H}_t \quad (13)$$

藉由平均向量法、權重式平均向量法、注意力法或自注意力法等四種模型架構，我們可以得到句子層次的向量表示法z。基於這個句子層次的表示法，我們採用兩層的前饋神經網路，進行最終的語音情緒辨識。

4.3 評估指標

在情緒辨識的任務中，我們首先使用平均倒數排名（Mean Reciprocal Rank, MRR），來呈現情緒辨識結果中，正確答案平均出現的位置。此外，我們也使用 Macro Average F1-score（Macro F1-Score）和 Weighted Average F1-score（Weighted F1-Score）作為評估指標。這兩者是常見的分類模型評估方法。Macro F1-Score 是對所有類別的 F1-score 進行平均，每個類別的 F1-score 在計算中具有相同的權重，因此不會因類別不平衡而偏向於較大的類別。Weighted F1-Score 則根據每個類別的樣本數量進行加權平均，更能反映模型在實際應用場景中的總體性能。總結來說，Macro F1-Score 更關注模型在不同類別上的一致性表現，而 Weighted F1-Score 則更關注模型在實際應用中的整體效能。

4.4 訓練與模型參數設置

我們使用 PyTorch 版本 24.05 實現本次的任務。每個情緒指標都視為獨立的任務，個別建立

單獨的模型。CEBE 語料庫長度最長為 11 秒，根據公式將每個音檔切成 11 塊，每一個區塊長度為 1 秒。模型訓練時，使用 Adam Optimizer，Batch 大小設定為 128，訓練週期（Epoch）設定為 100。我們採用 Focal Loss（T. -Y. Lin et al., 2020）作為損失函數，並根據發展集（Development Set）上的損失作為提前停止（Early Stopping）的判斷標準來保存最佳模型。此外，由於 CEBE 資料集情緒類別繁多，但每個類別所包含的數量相當不平均，因此我們將原先的 48 種情緒類別根據語意相似性重新歸類為十種情緒類別，如表 1 所示。

5 實驗結果與分析

使用 CEBE 資料集來進行語音情緒辨識的實驗結果如表 3~5 所示。在實驗中，我們比較了 8 種常見的聲學情緒特徵，並搭配本研究採用的 4 種句子層級之向量表示法模型。藉由比較表 3~5，有許多值得探討的結果。首先，從表 3 來看，不同的句子層級之向量表示法模型搭配不同的語音特徵集，都可以穩定地達到 0.8 左右的平均倒數排名分數。這顯示，正確的情緒類別大概都是預測結果中，機率第一或第二高的。但當我們更仔細地觀察表 3 的結果可以發現，不同的句子層級之向量表示法模型，需要搭配不同的特徵集，才能獲得最佳的結果：平均向量法＋IS2010、權重式平均向量法+IS2012、注意力法+emobase2010、自注意力法+emobase2010。這個結果使我們驚訝地發現，較複雜的特徵集並非一定能帶來較好的辨識結果，反而簡單的特徵集（IS2010）搭配平均向量法，就可以獲得很好的成績；此外，我們也觀察到，注意力法與自注意力法，並沒有在所有情況下都比平均向量法和權重式平均向量法獲得更好的結果。這可能是因為動態區塊分割模塊提供了一個很好的語音切割處理，不僅讓不同長度的語音句子可以有相同數量的區塊，也讓使得後續的模型僅需簡單的架構，就可以獲得良好的任務成效！當我們參考表 4 與表 5 的結果，發現在使用 F1-score 的計算下，平均向量法與權重式平均向量法更凸顯了與注意力法和自注意力法的差異。並且，因為 Recall 值較低的關係，所以 Macro F1-Score 的分數都很低。

6 結論與未來展望

本研究提出一個全新的中文教育廣播情緒語料庫（CEBE），為語音情緒識別領域帶來了新的挑戰與研究方向。雖然從我們的實驗中可以發現，平均倒數排名分數已可獲得不錯的結果，但在 Macro F1-Score 的評分上，結果則不理想。這些結果顯示，CEBE 資料集提供了一個大量且嶄新的平台，並且揭示了教育類語音資料在情緒識別任務中的獨特挑戰。教育語音在情感表達上的特殊性，應被仔細的思考與研究。

在未來，我們將更進一步地檢視、清理與再確認 CEBE 資料集的標籤正確性；探討更多現今卓越的語音情緒模型在 CEBE 資料集上的成效；探究教育語音與其他不同種類的語音內容之情緒表達的差異；著眼於為教育類語音情緒辨識提出一套新穎且有效的模型方法與訓練技術，從而提升教育類語音情緒辨識的準確性和泛化能力。

Acknowledgement

This work was supported by the National Science and Technology Council of Taiwan under Grants NSTC 113-2410-H-011-001, NSTC 112-2628-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

Alan S. Cowen, Dacher Keltner. Semantic Space Theory: A Computational Approach to Emotion. In *Cognitive Sciences, Volume 25, Issue 2*, Pages 124-136. 2021.

Björn Schuller, S. Steidl, and Anton Batliner. The Interspeech 2009 emotion challenge. In *Interspeech (2009)*, ISCA, Brighton, UK. 2009.

Björn Schuller, Anton Batliner, Stefan Steidl, Florian Schiel, Jarek Krajewski. The INTERSPEECH 2011 Speaker State Challenge. In *Proc. INTERSPEECH 2011*, ISCA, Florence, Italy, pp. 3201-3204, 28.-31.08. 2011.

Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, Benjamin Weiss. The INTERSPEECH 2012 Speaker Trait Challenge. In *Proc. INTERSPEECH 2012*, ISCA, Portland, OR, USA, 09.-13.09. 2012.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, Samuel Kim. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. Interspeech 2013*, pp. 148–152. Aug. 2013.

Burkhardt, Felix & Paeschke, Astrid & Rolfes, M. & Sendlmeier, Walter & Weiss, Benjamin. A database of German emotional speech. In *9th European Conference on Speech Communication and Technology*. 5. 1517-1520. 2005.

C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. In *Journal of Language Resources and Evaluation,* vol. 42, no. 4, pp. 335-359. December 2008.

F. Eyben, M. Wollmer, and B. Schuller. OpenSMILE: The Munich versatile and fast open-source audio feature extractor. In *Proc. ACM Int. Conf. Multimedia*, pp. 1459–1462. Oct. 2010.

Galen Andrew and Jianfeng Gao. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40. 2007.

Hume AI. Hume AI. https://www.hume.ai/. Accessed 10 Aug. 2024.

Jianfeng Zhao, Xia Mao, Lijiang Chen. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. In *Biomedical Signal Processing and Control*, Volume 47, Pages 312-323. 2019.

Microsoft. Speech AI Services. Azure. https://azure.microsoft.com/en-us/products/ai-services/ai-speech. Accessed 10 Aug. 2024.

R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. In *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37. 2010.

Russell, James & Mehrabian, Albert. Evidence for a Three-Factor Theory of Emotions. In *Journal of Research in Personality*. 11. 273-294. 1977.

S. Casale, A. Russo, G. Scebba and S. Serrano. Speech Emotion Classification Using Machine Learning Algorithms. In *IEEE International Conference on Semantic Computing*. Santa Clara, CA, 158-165. 2008.

Schuller, Björn & Steidl, Stefan & Batliner, Anton & Burkhardt, Felix & Devillers, Laurence & Müller, Christian & Narayanan, Shrikanth. The INTERSPEECH 2010 paralinguistic challenge. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*. 2794-2797. 2010.

T. Luong, H. Pham, and C. Manning. Effective approaches to attention-based neural machine translation. In *Proc. Conf. Empirical Methods Nat. Lang. Process.*, pp. 1412–1421. Sep. 2015.

T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár. Focal Loss for Dense Object Detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327. 1 Feb. 2020.

W. -C. Lin and C. Busso. Chunk-Level Speech Emotion Recognition: A General Framework of Sequence-to-One Dynamic Temporal Modeling. In *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1215-1227. 1 April-June 2023.

Y. Wang, A. Boumadane, and A. Heba. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*. 2021.

FineWeb-zhtw: Scalable Curation of Traditional Chinese Text Data from the Web

Cheng-Wei Lin 林承緯 [1*]; Wan-Hsuan Hsieh 謝宛軒 [1*]; Kai-Xin Guan 關凱欣 [1*]; Chan-Jan Hsu 許湛然 [1], Chia-Chen Kuo 郭嘉真 [2], Chuan-Lin Lai 賴傳霖 [2], Chung-Wei Chung 鐘崇偉 [2], Ming-Jen Wang 王明仁 [2], Da-Shan Shiu 許大山 [1]

[1]MediaTek Research [2]National Applied Research Labatories
{illiamw81, teresa920626, fionag1218}@gmail.com
{cckuo, c00lcl00, 2403047, c00mrw00}@narlabs.org.tw

Abstract

The quality and size of a pretraining dataset significantly influence the performance of large language models (LLMs). While there have been numerous efforts in the curation of such a dataset for English users, there is a relative lack of similar initiatives for Traditional Chinese. Building upon this foundation of FineWeb (Penedo et al., 2024a), we introduce FineWeb-zhtw, a dataset tailored specifically for Traditional Chinese users. We came up with multiple stages of meticulously designed filters to cater to the linguistic difference between English and Traditional Chinese, to ensure comprehensiveness and quality. We determined effectiveness from querying dataset samples with three main objectives. Our code and datasets are publicly available.

1 Introduction

The performance of large language models (LLMs) is significantly influenced by the quality and size of the pretraining datasets. While the exact data recipe remains undisclosed for many chatbot pursuits, it is generally perceived that trillions of high-quality tokens are the cornerstone for creating a generalist language model (Touvron et al., 2023a,b; Meta, 2024; Jiang et al., 2024). The vast expanse of the World Wide Web offers an enormous and diverse reservoir of data, making it an invaluable resource for experiments that require large-scale data. To extract relevant text information from raw HTML content, various preprocessing techniques have been established (Penedo et al., 2023b; Raffel et al., 2023; Gao et al., 2020). Among these methods, FineWeb (Penedo et al., 2024a) has set a new paradigm in this domain, boasting a very competitive data-performance trade-off. Despite the many efforts in English, there remains a notable absence of publicly available, high-quality datasets specifically tailored for Traditional Chinese. Further, the unique linguistic and cultural nuances of Traditional Chinese, such as lack of spacing, hinder the direct projection of current pipelines, which necessitates a specialized refinement approach.

In response to this need, we introduce FineWeb-zhtw, a dataset meticulously curated to provide a robust and comprehensive collection of Traditional Chinese text catered to the language usage of the Taiwanese community. By employing advanced filtering techniques, we ensure high quality and relevance on the samples of FineWeb-zhtw.

The creation of FineWeb-zhtw involves several stages of filtering, including basic filtering, multi-staged language identification, followed by Gopher (Rae et al., 2022), C4 (Raffel et al., 2023), and FineWeb (Penedo et al., 2024a) quality filters. Each of these stages is designed to enhance the dataset's overall quality and relevance. For the implementation of these filters and the curation process, we utilized `datatrove` (Penedo et al., 2024b), a tool developed by the HuggingFace team that facilitated our efforts effectively. We then evaluate the dataset samples on three custom metrics to understand the effectiveness of our filtering pipeline.

Through FineWeb-zhtw, we aim to advance the development of natural language research for Traditional Chinese, providing a crucial re-

* Equal Contribution. Work done during internship at MediaTek Research. 林承緯 is studying at the Department of Computer Science of National Yang Ming Chiao Tung University, 謝宛軒 and 關凱欣 are studying at the Department of Information Management of National Taiwan University

source for the community to build upon.

2 Building FineWeb-zhtw

The raw data for FineWeb-zhtw is sourced from Common Crawl, an extensive web archive that captures a comprehensive snapshot of the internet. The data is stored in Web ARChive (WARC) format, which is a standard format for web-crawled data. This format includes both the retrieved content and metadata, making it suitable for large-scale data processing tasks. For the parameters mentioned in the paragraph, we determined the parameters using grid search by manually inspecting the quality of filtered-in and filtered-out data.

2.1 Basic filtering and Text Extraction

The initial stage of building the FineWeb-TC revolves around extracting text from html webpages. We use the `trafilatura` tool favoring precision to remove potential headers and footers. Since html cleaning is an expensive operation compared to filtering out entire webpages, we implemented a pre-filter that excludes documents that are guaranteed not to be in Traditional Chinese to minimize computational cost. The pre-filter defines fuzzy Traditional Chinese tokens by an unicode range[1], and any document without 5 consecutive characters in the set will be filtered out.

We also applied a URL filter, which blocks flagged websites that have low content value such as spam.

2.2 Language Identification

The filtering process for FineWeb-TC is to identify core text from html that are naturally written in Traditional Chinese. Most language identifiers, such as fasttext (Bojanowski et al., 2017), do not differentiate between Traditional Chinese and Simplified Chinese. As we could not find a pre-defined document-wise filter with satisfying accuracy on these two variants of Chinese, we designed a custom filter in addition to fasttext, with character and word-phrase filtering. We remove documents that match any of the specified phrases, prioritizing the precision of the remaining entries. This step is crucial for ensuring that the dataset accurately reflects and aligns with the language and preferences of the local community.

[1] $\backslash u3040 - \backslash u3090 \backslash u30a0 - \backslash u30ff \backslash u4e00 - \backslash u9fff$

2.3 Gopher filters

Next, we applied the Gopher Quality Filter, which uses heuristic rules to ensure the text content adheres to specific quality standards. This custom filter evaluates various textual characteristics and excludes documents that do not meet the established criteria.

- **Document Length**: Short documents often lack useful information, while excessively long documents are more likely to be spam and can skew the dataset. Therefore, we excluded documents with fewer than 50 words or more than 100,000 words.

- **Symbol-to-Word Ratio**: Excessive use of symbols can indicate low-quality content. To filter these out, we excluded documents with a symbol-to-word ratio exceeding 0.1.

- **Ellipsis Lines Ratio**: High frequencies of ellipses can suggest truncated content. These articles with unfinished sentences decrease in value tremendously under the auto-regressive modeling objective. Documents with more than 30% of lines ending with ellipses were excluded.

- **Stop Words Count**: Documents lacking sufficient stop words often do not contain expressive and coherent sentences. Therefore, we excluded those without any of the predefined stop words.

By excluding documents that fail to meet these criteria, the Gopher Quality Filter ensures a dataset with more consistent and high-quality content suitable for analysis and application.

2.4 C4 filters

To further refine the FineWeb-zhtw dataset, we implemented the C4 Quality Filter and customized its parameters to optimize for filtering Traditional Chinese content. Besides document-wise filtering, the C4 filter also highlights quality checks on each line of the documents. Uninformative lines are removed to increase the density of relative information in the data.

The C4 Quality Filter uses several key criteria to determine the quality of a line in the content:

- **Filter JavaScript**: Lines containing the word `javascript` or curly brackets are often code snippets or irrelevant content. These lines are excluded to keep the dataset focused on textual content.
- **Filter Policy-Related Substrings**: Lines containing policy-related substrings such as `terms of use`, `privacy policy`, or `cookie policy` are usually boilerplate text. Excluding these lines helps remove non-informative content.
- **Bracket Ratio**: A high bracket ratio can indicate the presence of excessive code or other structured content. Documents with a bracket ratio exceeding 0.01 are excluded. The bracket ratio is calculated as the proportion of brackets in the document.

By applying these criteria, the FineWeb-zhtw dataset is cleaner, more relevant, and better suited for subsequent analysis and applications.

2.5 FineWeb quality filters

Finally, we implemented the FineWeb Quality Filter and tuned its parameters to enhance the quality of Traditional Chinese content. This custom filter ensures that the text content meets high-quality standards by analyzing key textual characteristics and excluding documents that do not meet these criteria.

The FineWeb Quality Filter uses the following criteria to assess document quality:

- **Line Punctuation Ratio**: Proper punctuation is essential for text coherence. We exclude documents with a line punctuation ratio below 0.04 to ensure that only well-punctuated, readable content is included.
- **Short Line Ratio**: Documents with a high ratio of lines shorter than 10 characters are likely fragmented. By excluding documents where this ratio exceeds 0.8, we ensure that the dataset contains substantial and informative content.
- **Character Duplication Ratio**: Excessive repetition of characters can indicate low-quality content. We exclude documents with a duplication ratio above 0.3 to filter out repetitive text and retain diverse content.
- **New Line Ratio**: A high new line ratio suggests poor text structure. We exclude documents with a new line ratio greater than 0.3 to maintain well-organized and coherent documents.

By applying these criteria, the amount of low-quality or poorly structured content in the dataset is significantly reduced.

2.6 Minhash Deduplication

We first applied a document-level minhash deduplication to remove repetitive content. Then, we iteratively removed lines of content leading or trailing the document, if that line has over 100 occurrences in one dump of Common Crawl.

2.7 The final FineWeb-zhtw dataset

We have implemented a comprehensive pipeline to process Common Crawl data, utilizing various filtering techniques at each stage. The process begins with extracting documents from WARC files. Then, filters mentioned in the sections above are applied sequentially, starting with the basic filtering (Section 2.1), followed by the Language Identification (Section 2.2). Next, we use the gopher filters (Section 2.3), C4 filters (Section 2.4), FineWeb quality filters (Section 2.5), and finally, minhash deduplication (Section 2.6).

Figure 1 illustrates the global removal rate and the relative retention rates of documents at each stage, offering not only an overview of the total Traditional Chinese data reserve but also a comparative analysis of the impact of each filter. The final screening rate is around 99.5%, leaving 0.5% of effective Traditional Chinese data available.[2]

[2]14.04 gigabytes of pure text data remain for the dump CC-MAIN-2024-26

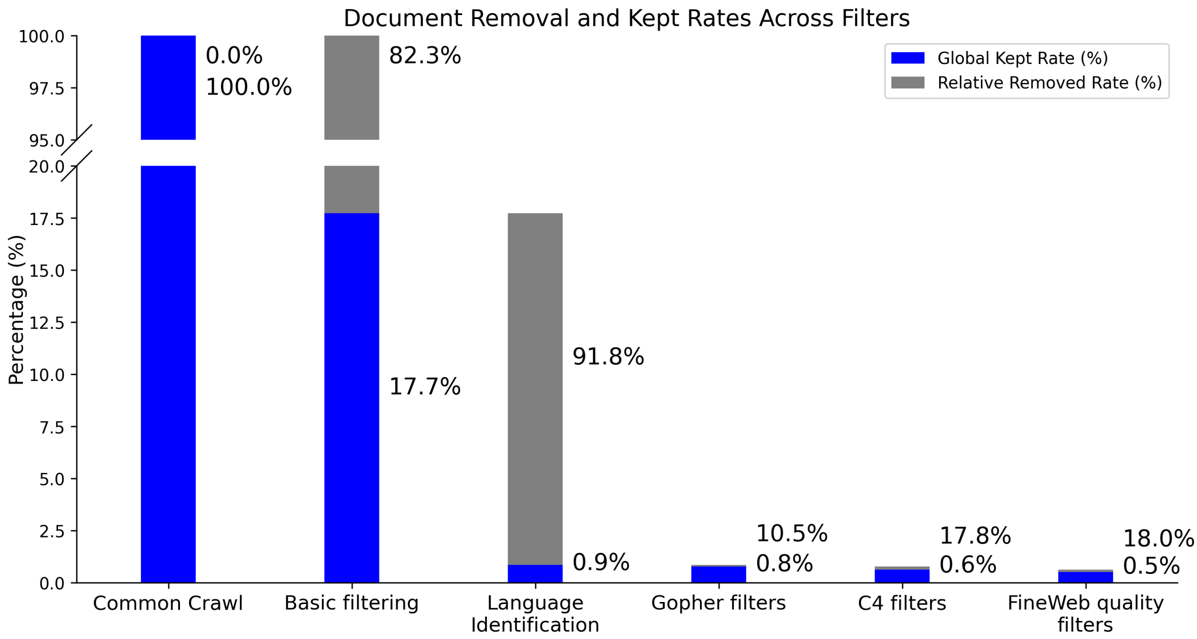


Figure 1: **Relative removal rate and global kept rate of Common Crawl during each filtering stage.** We report the removal rate with respect to each previous stage, and the overall kept rate (Penedo et al., 2023a). Rates are calculated by the number of documents until the language identification stage. For later line-by-line filtering, rates are measure in tokens (bytes).

3 Evaluation

3.1 Evaluation Methodology

To ensure the quality and relevance of the FineWeb-zhtw dataset, we evaluated the final dataset using the following criteria:

1. **Traditional Chinese and Language Naturalness**: Accurate use of Traditional Chinese characters and natural language is essential for accessibility and usability by native speakers. Proper language usage enhances readability and comprehension, which are critical for any linguistic dataset.

   This criterion assessed whether the content is written in Traditional Chinese, employs accurate characters, and adheres to Traditional Chinese grammar and conventions. Factors include grammatical correctness, character usage, contextual appropriateness, clarity, and logical structure.

2. **Educational Value**: High educational value ensures that the content contains relevant knowledge. Datasets with great educational content provide reliable materials that enhance knowledge-based question answering of the language model.

   This criterion evaluated whether the content is informative and relevant for learning purposes, considering educational coverage, coherence, structure, and overall value.

3. **Sensitive Content**: Excluding sensitive or inappropriate content is vital for ethical development and ensuring user safety.

   This criterion checked for sensitive or inappropriate content, including violence, pornography, discrimination, and privacy violations.

Each document was rated on a scale from 0 to 5 for each criterion, and the scores were summed to provide a total score for each document.

We utilized ”language model as a scoring agent” (Chiang and Lee, 2023; Chiang et al., 2024) to automate and streamline this evaluation. The assistant was given a detailed prompt outlining the criteria and scoring methodology (See Appendix for details), ensuring consistent and accurate assessment.

We randomly selected 1,000 documents from different stages of the pipeline and ran evaluations on the metrics mentioned above, using GPT3.5 on the DaVinci[3] platform.

3.2 Evaluation Results

To assess the effectiveness of our filtering process and the quality of the FineWeb-zhtw dataset, we compared it against two ablations: the dataset after basic filtering and the dataset after language identification.

Figure 2 presents the evaluation results for the different datasets, highlighting the distribution of scores and mean values for each criterion. We conducted a t-test on the null hypothesis (H_0) that the scores are identical. Based on the resulting t-statistic and p-values, we rejected the null hypothesis for p-values less than 0.05. Results show that the FineWeb-zhtw dataset has consistent improvements across all categories compared to the datasets after basic filtering and language identification.

In the **Traditional Chinese and Language Naturalness** category, FineWeb-zhtw led with a mean score of 2.42, compared to 2.19 for the dataset after language identification and 1.72 for basic filtering. Significant improvements with FineWeb-zhtw are supported by a t-statistic of 5.36 (p-value = 9.14e-08) and for the dataset after language identification over basic filtering by t-statistic = 10.13 (p-value = 1.53e-23). The significant difference between FineWeb-zhtw and basic filtering is highlighted by a t-statistic of 15.23 (p-value = 1.29e-49).

For the criterion **Educational Value**, FineWeb-zhtw scored a mean of 2.04, outperforming the dataset after language identification (1.77) and basic filtering (1.54). Significant differences are evident with FineWeb-zhtw compared to the dataset after language identification (t-statistic = 4.96, p-value = 7.56e-07) and the dataset after language identification compared to basic filtering (t-statistic = 4.12, p-value = 3.93e-05). The difference between FineWeb-zhtw and basic filtering is supported by a t-statistic of 8.78 (p-value = 3.38e-18).

In the **Sensitive Content** category, as the score is already high, no statistically signifi-

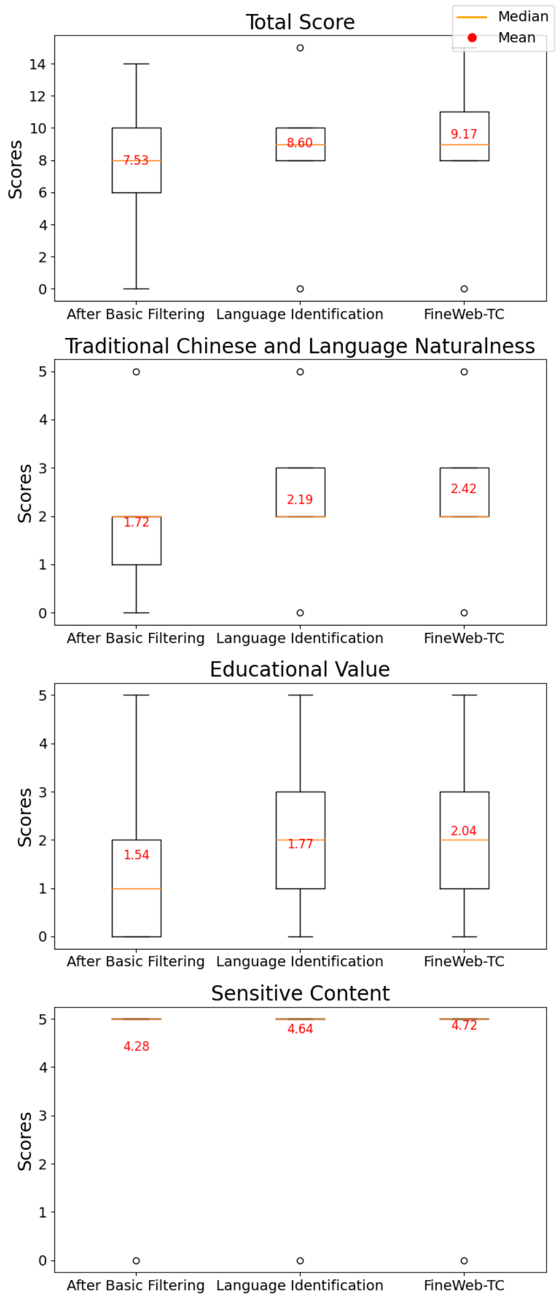


Figure 2: Comparison of evaluation results across different datasets. The FineWeb-zhtw dataset demonstrates significant improvements in Total Score, Traditional Chinese and Language Naturalness, and Educational Value compared to the dataset after language identification and basic filtering, with statistically significant differences (as indicated by t-test results).

[3]https://dvcbot.net/

cant gains are achieved from the FineWeb filtering stage. These scores reveal that sensitive content is not a huge concern when using CommonCrawl data under our pipeline.

Finally, for the **Total Score**, FineWeb-zhtw achieved a mean of 9.17, higher than 8.60 for the dataset after language identification and 7.53 after basic filtering. T-tests confirm significant differences between all pairings.

These results highlight that our sophisticated filtering pipeline is effective at extracting crucial assets that have a higher educational value and language coherence, which have shown to benefit language model training (Penedo et al., 2024a).

4 Discussion

Through the realization of FineWeb-zhtw, we have also observed a significant gap between readily available English and Traditional Chinese data. At the language identification stage, the gap between the data size of English and Traditional Chinese is already around 40x. From our empirical screening rate, it is estimated that even if the entire Common Crawl data is completely extracted to train a Traditional Chinese LLM with 70B parameters, the amount would not be sufficient under the Chinchilla scaling law (Hoffmann et al., 2022). It is therefore of great importance that additional efforts be made to curate publicly available data beyond what is found in Common Crawl.

5 Conclusion

In summary, the FineWeb-TC dataset, resulting from the advanced filtering process, represents a significant advancement in the field of readily available Traditional Chinese language modeling corpus. The advanced techniques applied have successfully improved the dataset's overall quality, ensuring better adherence to language standards, educational value, and content appropriateness. These findings affirm the efficacy of our comprehensive filtering pipeline in producing a high-quality Traditional Chinese dataset. Finally, we highlight the qualitative differences between English and Traditional Chinese FineWeb that arise from quantitative changes.

Acknowledgements

We appreciate the accelerative enhancement made by the members of the project titled 'Smart Decision-Making Mobile Platform for Social Workers' (No. M1304129) in the study design.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung yi Lee. 2024. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Meta. 2024. Meta-llama-3.1-8b-instruct.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin

Raffel, Leandro Von Werra, and Thomas Wolf. 2024a. The fineweb datasets: Decanting the web for the finest text data at scale.

Guilherme Penedo, Hynek Kydlíček, Alessandro Cappelli, Mario Sasko, and Thomas Wolf. 2024b. Datatrove: large scale data processing.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023a. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, 36:79155–79172.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023b. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis insights from training gopher.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Appendix

Evaluation Details

We show here the prompt that we use for scoring the data samples:

評量標準

1. 繁體中文與語言自然性：內容是否以繁體中文寫成，並使用正確的繁體中文字符；文本是否自然流暢，符合繁體中文的語法和用詞習慣，並且易於閱讀和理解、內容完整。滿分 5 分，評分時請考慮以下因素：
    - 若語法正確，句子結構自然，得 1 分。
    - 若使用正確的繁體中文字符，得 1 分。
    - 若用詞符合繁體中文的習慣，得 1 分。
    - 若句子簡潔明瞭，易於理解，得 1 分。
    - 若內容有邏輯性，有頭有尾，得

1 分。

2. 教育價值：內容是否具有值得學習的正面價值。滿分 5 分，評分時請考慮以下因素：
   - 若內容提供與教育話題相關的基本資訊，即使其中包含一些無關或非學術的內容（如廣告和促銷），得 1 分。
   - 若內容涉及某些教育相關元素，但不完全符合教育標準，得 1 分。它可能會混合教育內容和非教育教材，提供潛在有用主題的概述，或者以無條理和不連貫的寫作風格呈現訊息。
   - 若內容適用於教育用途並介紹與學校課程相關的關鍵概念，得 1 分。它是連貫的，儘管可能不全面或包含一些無關的內容。它可能類似於教科書的介紹部分或適合學習但有顯著局限性的基礎教程，比如將概念處理得對於初中學生來說過於複雜。
   - 若內容對於小學或初中等級的教育目的是高度相關且有益的，得 1 分。它展現出清晰和一致的寫作風格，可能類似於教科書的一章或教程，提供大量的教育內容，包括練習和解答，且幾乎沒有無關的內容，概念不會太先進而超出初中學生的理解範圍。內容連貫、集中且有價值的結構化學習。
   - 若內容在教育價值上是出類拔萃的，完全適合在小學或初中教學，得 1 分。它遵循詳細的推理，寫作風格易於理解，提供深入而透徹的主題見解，且沒有任何非教育或複雜內容。

3. 敏感內容：是否包含敏感或不適當的內容。滿分 5 分，評分時請考慮以下因素：
   - 若不包含暴力言論、行爲、與內容，不宣揚言語、肢體等任一形式暴力，不宣揚槍枝與血腥等內容，得 1 分。
   - 若不包含色情言論、行爲、與內容，得 1 分。
   - 若不包含歧視言論、行爲、與內容，不貶低、侮辱或仇恨任一種族、國家、族群、與個人，得 1 分。
   - 若不涉及政治和宗教等敏感話題，不以言論、行爲等任一方式支持或反對任一政黨或教派，得 1 分。
   - 若不包含侵犯隱私或個人權利的言論、行爲、與內容，得 1 分。

評分格式

請按照以下格式提供評分，分數應以整數型態表示，並將每個標準所得分數加總，計算總分：

1. 繁體中文與語言自然性：< 分數 >
2. 教育價值：< 分數 >
3. 敏感內容：< 分數 >

總分：< 總分 >

評分範例

請對以下文本進行評分：
< 待評估的文本 >

Design of a mSUD Taiwan Taigi Treebank Aligned on Mandarin and Teochew Translations

Pierre Magistry and **Ilaine Wang**
ERTIM – Inalco
2, rue de Lille
75007 Paris, FRANCE
pierre.magistry@inalco.fr

Chen Siman 陈思漫 | **Chen Xinlei** 陈新蕾 | **Li Zhongjie** 李仲杰 | **Wen Yu-Chieh** 温侑洁 | **Zhang Weiqi** 张炜祺

Abstract

This paper presents the design choices and the first (preliminary) release of a trilingual treebank of Taigi sentences aligned on translations into Mandarin and Teochew. The three languages come with morphosyntactic annotations following the joint morphology and Surface syntactic Universal Dependencies (mSUD) scheme. We provide 54 annotated and validated sentences as a first release for open discussion, with the objective of annotating the full set of 420 sentences from the examples in Taiwan Ministry of Education's Dictionary of Frequently-Used Taiwan Minnan.

Keywords: Treebank, Syntax, Morphology, Taigi, Mandarin, Teochew, mSurface Syntactic Universal Dependency

1 Introduction

This paper presents a syntactic treebank for Taiwan Taigi (hereafter "Taigi") aligned with two other Sinitic languages: Taiwan Mandarin and Teochew. Our motivation is to provide manually annotated data which will allow for linguistic description, cross linguistic comparisons and evaluation data needed to describe more precisely the knowledge encapsulated into various large language models.

We chose to follow the Universal Dependencies (UD) framework to ease alignment and cross linguistic comparisons. More specifically, we adopt the mSUD scheme to provide a description in terms of surface syntax (Guillaume et al., 2024) and a morph-level annotation to tackle the issues related to word segmentation more precisely.

2 Languages of the Treebank

Taigi is a Sinitic language spoken in Taiwan, with a rich history and a complex sociolinguistic situation. Taigi is closely related to other Southern Min languages spoken in the South of China and in numerous diasporan communities in South East Asia, but it also has unique properties resulting from the history of Taiwan, especially language contacts with Austronesian languages and Japanese. It is the most widely used language in Taiwan after Mandarin. After being prohibited until the late 20th century, it is now receiving political support and is undergoing the process of standardization by the Ministry of Education (MOE) in Taiwan.

Our dataset is based on an official release of a set of example sentences contained in the Dictionary edited by the MOE and published as Open Data under a Creative Commons license[1]. The text in Taigi is provided both in sinograms (commonly called *Chinese characters*) and in Tâi-lô romanization. The dictionary also includes translation in Mandarin for every sentence.

The Mandarin in this corpus (hereafter MSM, for Modern Standard Mandarin) corresponds to the official variant of Mandarin used by the Republic of China (ROC, Taiwan). It may differ from MSM in use in China, especially in terms of lexical choices, but also for some syntactic constructions. Taiwanese MSM is more subject to the influence from Taigi and this characteristic is probably slightly amplified by the translation nature of this specific corpus in which the text is originally written in Taigi.

[1] https://sutian.moe.edu.tw/zh-hant/siongkuantsuguan/

Teochew is a language originally spoken in the extreme South East of the Guangdong province in China, next to the Fujian province border. It is thus expected to be quite distant from Taigi but closer than MSM. Teochew is also spoken in many places by diasporan communities and is subject to variation from one community to another. This work is part of a broader project based in Paris where one of those communities lives, allowing us to work closely with Teochew people. Incidentally, this study contributes to a larger endeavor to characterize Teochew varieties. This will enable us to draw a more comprehensive picture, make more detailed comparisons and conduct transfer learning experiments.

According to our own observations of native speakers, Taigi and Teochew nevertheless allow for a certain degree of mutual intelligibility, but requiring noticeable efforts from both sides (and maybe a sense of Sinitic languages phonology). One of the participants of this project is a native speaker of Teochew from China who was in charge of providing translations (see below for more details). In contrast with Taigi, Teochew is not ongoing a standardization process. The choice of scripts is therefore less straightforward in this case, as we will explain in the next section.

2.1 Scripts of the Treebank

The languages of our corpus can be written using a variety of scripts.

sinograms The three languages can be written in sinograms. Following the original dataset, we use traditional variants for Mandarin and the official characters advocated by the MOE for Taigi without any modification of the original data. Selecting sinograms for Teochew is less easy as there is no clear official recommendation. For the sake of consistency we avoid simplified characters and limit our choice to traditional characters. This may be unusual for Teochew in China, but this practice is also attested in diasporan communities, so this choice is not too specific to our treebank. When unsure, we rely on two dictionaries.[2]

romanizations The three languages can also be written using the Latin script. MSM is typically romanized with *hànyú pīnyīn*, but the transcription was not included in the data released by the MOE. Regarding Teochew, the *Guangdong Peng'im* has been designed in 1960 by the Guangdong province government, and later slightly modified in the diasporan communities and turn into *Gaginang Peng'im* which is more suited for US keyboards. Converting between the two systems is straightforward using formal grammars[3] so we plan to release the treebank in both versions. Examples given in this paper are romanized with Guangdong Peng'im. Different romanizations are possible for Taigi. The most widely used are the historical *Peh-oe-ji* and the now official Tâi-lô. The use of the latter is advocated for by Taiwan's Ministry of Education and the dataset we use includes this romanization. Although conversion between the two system can also be achieved fully automatically if needed, we chose to follow the recommendation and use Tâi-lô.

2.2 Organization of the Work

This work is part of a project which aims to study the limitations of current methods and tools in Natural Language Processing (NLP) when used for digital humanities research on Sinitic language resources. For this purpose, we first need resources covering the three axes of variation we wish to study: diachrony, diatopy, and grapholinguistics.

This specific part on the treebanks was conducted by a team of 5 MA students in NLP, advised by two researchers. The students have a variety of linguistic background which greatly contributed to the success of this work. All the students are proficient in MSM (either Taiwan Mandarin or China *putonghua*). One was also a native speaker of Taigi from Kaohsiung, another a speaker of Amoy Hokkien, and another a native speaker of Teochew from China. The two researcher have experience in Sinitic languages processing and linguistics. One has a decent command of Mandarin and knowledge

[2] http://www.czyzd.com/ and https://play.google.com/store/apps/details?id=com.tcknow.whattcsay3&hl=en_AU&gl=US

[3] A conversion tool was already available here: https://github.com/learn-teochew/parsetc

of Taigi grammar, while the other is an heritage speaker of the Teochew diaspora.

Students worked in pairs, issues were discussed during regular meeting and all the final trees are reviewed by one researcher.

For the Teochew part, the native speaker was asked to do the translation from the MSM version to limit the influence of Taigi in her lexical and grammatical choices.

3 Related Works

Despite increasing efforts to build corpora for Taigi, recent works focus more on large amount of raw data to train language models and speech corpora (Liao et al., 2020). Resources with morphosyntactic annotations for Taigi are still very scarce. Noticeable exceptions include the early T3 corpus (Chou, 2006) which was not publicly released, and an attempt to build a Part of Speech (POS) tagged corpus semi-automatically, based on Mandarin resources (Iunn et al., 2009). Tsay (2007) describes what appears to be the only corpus manually annotated in morphosyntax, but it is from a very different genre (spontaneous child speech) and includes only POS tagging (no dependency syntax). To our knowledge, our treebank is thus the first of its kind.

On the other hand, various syntactic treebanks for MSM are available. The UD website contains no less than 6 different corpora for MSM. Two of them come with a mSUD version: Chinese Beginner[4] and Chinese PatentChar[5]. Our work starts from their guidelines and we do our best to stay compatible in terms of linguistic descriptions. Another important related treebank is the UD conversion of the Sinica Treebank (Hsieh et al., 2022), which is closer to the MSM in our corpus, but it only uses the UD scheme and was converted from the original Sinica Treebank (Huang et al., 2000) with a rule-based system that introduces a small amount of errors.

Universal Dependencies (Nivre et al., 2020) is a well known project which aims at building multilingual treebanks following consistent guidelines across different languages, to ease NLP and linguistic comparison. The Surface-Syntactic Universal Dependencies (SUD) (Gerdes et al., 2018) is an annotation scheme which follows distributional criteria closer to more traditional dependency syntax theories than UD (favoring syntactic rather than semantic heads). It still aims at being fully compatible with UD guidelines through automatic conversion using graph rewriting rules.

The mSUD scheme (Guillaume et al., 2024) is an extension to SUD allowing for the joint annotation of syntax and morphology in the same formalism by using morphs as leaves of the dependency trees. It has been used for MSM in Li et al. (2019). Similar discussions can be found in the literature of Chinese processing advocating for character-level parsing (independently from the UD project) such as Zhang et al. (2014, 2013). Conversion from mSUD to word-level SUD is also straightforward with graph-rewriting tools.

Our work also relies on the tooling offered by the UD and SUD community, namely Arborator-Grew (Guibon et al., 2020) for online annotation and Grew (Guillaume, 2021) for graph queries and transformations.

4 Annotation Layers

4.1 Word Segmentation

Word Segmentation is an expected issue when addressing textual annotation for Sinitic languages. A typical treebanking workflow for such languages written in *scriptio continua* is to start with defining segmentation guidelines as a prerequisite before starting the syntactic analysis. But word segmentation decisions and syntactic analysis are closely related, and it is more convenient to address both jointly. We do so by adopting the mSUD scheme which allows us to somehow delay the word segmentation analysis while conducting the full syntactic analysis of each sentence.

Another helpful feature of this Taigi corpus is that it comes with a Tâi-lô romanization. Taigi has a long history of digraphia, being written either in sinograms for centuries or romanized since the end of the 19th century. The use of the Latin script has long in-

[4] https://universaldependencies.org/treebanks/zh_beginner/index.html

[5] https://universaldependencies.org/treebanks/zh_patentchar/

troduced the use of a non *scriptio continua* and a word-level writing. This script used to be more widely adopted and has been an actual writing system for publication (contrary to hànyú pīnyīn and 注音符號 zhùyīn fúhào for Taiwan MSM which are only used as phonetic transcription, mostly for educational purposes). As a result, we can rely on the Tâi-lô text to provide a first good approximation of a word-level tokenization.

Just like any language written with the Latin script and spaces, the correspondence between orthographic words and syntactic units is not perfect – typically when dealing with multiwords expressions.

Segmentation is discussed in the appendix of the guidelines for official Taigi romanization followed by the editors of the dictionary[6], especially some ambiguous cases where different segmentations correspond to different meanings. Such examples include *âng-hue* vs. *âng hue* (sinograms: 紅花, lit. red+flower), which correspond respectively to a medicine plant name (non compositional meaning) or to a red flower. In our treebank this would result in different morphosyntactic relations, resp. `/m` or `mod`. The same goes for *oo niau* vs. *oo-niau*, (sinograms 烏貓) resp. a black cat or a fashionable girl. Longer frozen expressions also exist such as *tshenn-mê-gû* 青盲牛 for illiterate, non compositionnal meaning from *thsenn-mê* '*blind*' and *gû* '*cow*'[7].

Such ambiguities can also concern other relations and affect parts of speech, for example 風吹 (lit. *wind+blowing*) may mean *the wind blows* with a `subj` relation, in which case it is written *hong tshue*, or it may mean *a kite* (non compositional meaning) in which case it is written *hong-tshue* and annotated with a `/m` relation.

In the case of our Taigi treebank, specific structures such as Verb-Object constructions can lead us to introduce some discrepancies with the Tâi-lô version as illustrated in Figure 1. This confirms the benefits of using the mSUD scheme even when annotating corpora in Latin script.

4.2 Part of Speech Tags

One major difficulty to tackle when building a new UD treebank is to follow their standard POS tagset. For each treebank, the annotation guidelines must detail how the limited and fixed tagset of only 17 tags[8] was adapted.

In our case, the most striking example of such difficulties is the lack of a specific tag for nominal classifiers. Therefore, we decided to follow the common practice for MSM treebanks to annotate classifiers as a nominal elements (`NOUN` tag) and to mark the classifier function as a `clf` syntactic relation with the head noun. Another issue is the question of the adjectives and the distinction between stative and action verbs. As we only have a single `VERB` tag for all subtypes of verbs, we use the `ADJ` tag for predicative adjectives and keep `VERB` for action verbs.

More generally, we adopt the following strategy:

1. We first try to follow the same guidelines as the MSM corpora annotated in mSUD[9]. Most of the discussions are relevant for Sinitic languages with only little adaptation for Taigi.

2. We discuss unclear cases during meetings or through github issues.

3. We check categories used in the MOE's Dictionary of Frequently-Used Taiwan Taigi[10]

An interesting issue arises from the differences of tagsets between UD and the MOE dictionary. The latter adopts the categories from traditional Chinese lexicography. We did not try to define a systematic mapping between the two tagsets as we expect difficult cases of ambiguities. However, it is interesting to observe the correspondences resulting from our annotation afterwards.

We project the categories from the dictionary on our corpus by assigning the list of possible traditional categories to the `xpos`

[6] https://language.moe.gov.tw/001/Upload/FileUpload/3677-15601/Documents/tshiutsheh_1081017.pdf p.55

[7] in such case syntactic structure may change, a '*blind cow*' would more likely be translated as *tshenn-mê ê gû* 青盲的牛.

[8] https://universaldependencies.org/u/pos/all.html

[9] https://guidelines.surfacesyntacticud.org/docs/language/mandarin_chinese/

[10] https://sutian.moe.edu.tw/zh-hant/

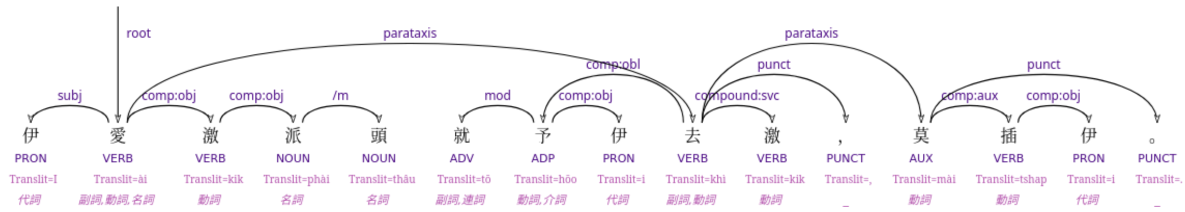


Figure 1: *He likes to put on airs, just let him put on airs and ignore him.*
Example of discrepancy between syntactic units and orthographic word segmentation. Here *kik-phài-thâu* is written in a single word (meaning *to put on airs*, lit. *to arouse, stimulate* (kik) *a dignified air* (phài-thâu), but despite its non compositional meaning, the Verb-Object structure is clear and required to explain the repetition of *kik* without *phài-thâu*.

property of each tokens. In order to do so, we build the words from the graph by extracting subtrees connected by morphological (`/m`) relations and try to match the sinograms and pronunciation to a dictionary entry (See Figure 1 for an example of a fully annotated sentence). Since the corpus is based on dictionary examples, the coverage is almost perfect. On the other hand, we do not try to resolve the ambiguities and simply record the list of possible categories for each word. Subword tokens are assigned the `xpos` of the word they appear into.

This allows us to draw the following Sankey diagrams. Such illustration is helpful not only to understand and document our annotation choices, but also to compare traditional categories and UD tagset and for error mining, combined with the power of ArboratorGrew query language.

Figure 2 is based on the whole corpus and contains both clear cases of good correspondences at the bottom, such as 代詞 – `PRON`, 副詞 – `ADV`, 動詞 – `VERB`, 名詞 – `NOUN`, 形容詞 – `ADJ`, 助詞 – `PART` and more complex cases which deserve more discussion are shown.

Figure 3 is based on the same data, with the aforementioned correspondences removed to focus on the less obvious cases.

The following observations are examples of what we can infer from the two figures:

- for Sinitic languages, we expect a high level of lexical ambiguity as 0-derivation is often possible to change the POS of a word, but we also see that a large part of the vocabulary is actually very stable;
- the idioms (熟語) in the dictionary are a subclass of verbal expressions;
- position words (方位詞) are split into a subclass of nouns and a subclass of adverbs (with possible lexical ambiguity);
- prepositions (介詞) are split between actual prepositions (`ADP`) and `VERB`s. This is to be expected as 介詞 are usually grammaticalized verbs, resulting in lexical ambiguity;
- as explained earlier, we treat classifiers (量詞) as a special case of nominal items and tag them as `NOUN`;
- temporal expressions (時間詞) are mostly nominal expressions, aside from an occasional adverbial usage;
- 助詞, some kind of particles seems to cover not only what we tag `PART`, but also encompass different kinds of items. It is likely to be more related to actual functions of the word than it is to POS which definitely deserves more investigation;
- our class of `AUX` (auxiliary) is clearly a subclass of verbs, but with many lexical ambiguities. Some are also classified as 副詞 (closer to adverbs) in traditional analysis, but we consider auxiliaries as syntactic heads, so the `ADV` tag would not be appropriate. In UD, `AUX` must be a closed list of items, so further investigation is required to strictly define this list based on our preliminary annotations.
- the 代詞 category is often translated as “pronoun”, but actually includes determiners.

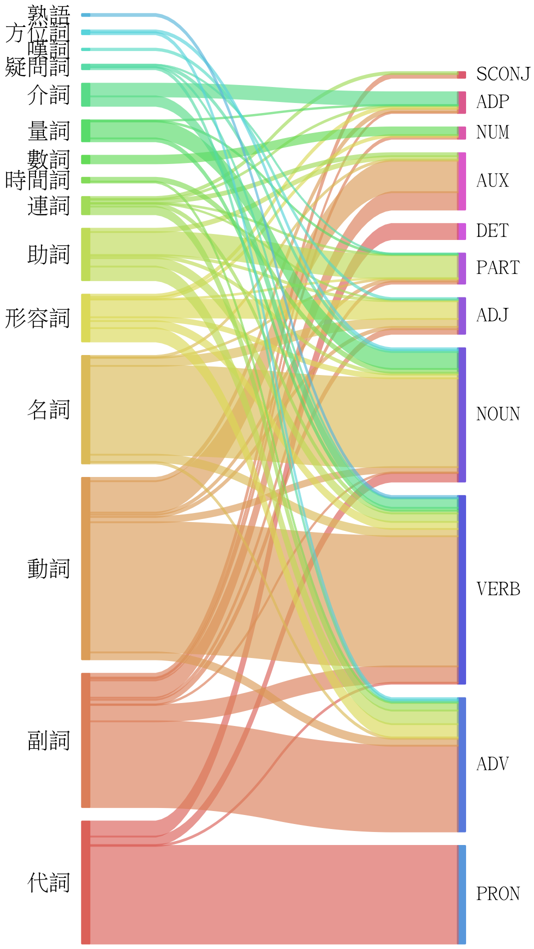


Figure 2: Sankey diagram to show correspondences and mismatches between the (S)UD POS tagset and categories from the MOE Dictionary.

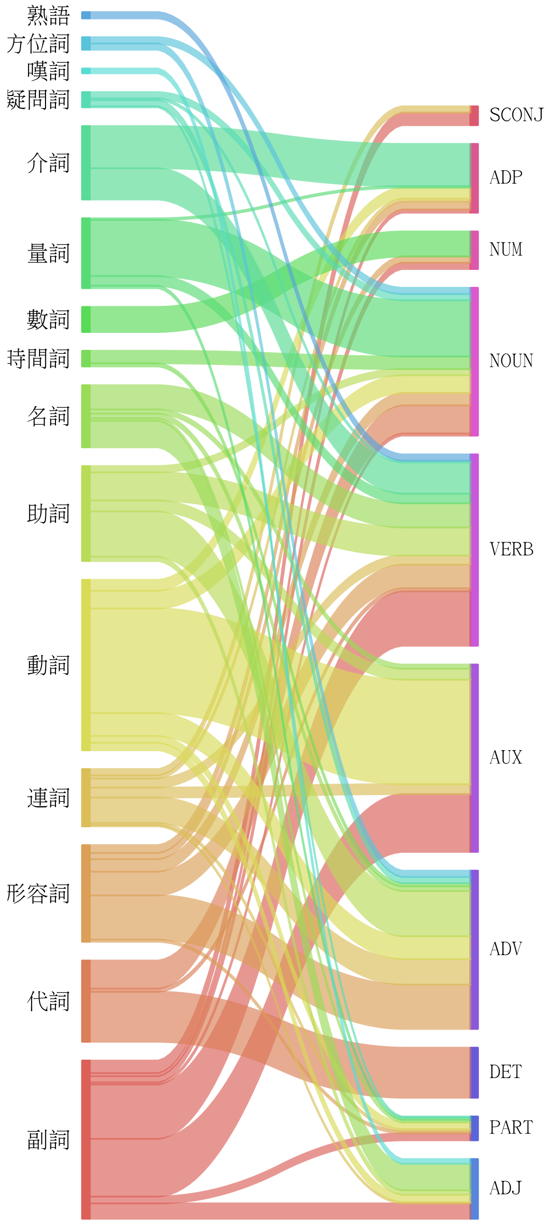


Figure 3: Sankey diagram to show correspondences and mismatches between the (S)UD POS tagset and categories from the MOE Dictionary, with naively expected mapping removed.

4.3 Syntactic Relations

We used the following syntactic relation (sorted by frequency):

mod for modifiers
punct for punctuation marks
comp:obj for verb – (direct) object relations
subj for verb – subjects relations
comp:aux for complements of auxiliary verbs
parataxis between heads of clauses without explicit syntactic or discourse connector
compl:obl for oblique (indirect) objects
discourse:sp for sentence final particles
det for noun phrase determiners
comp:res for resultative constructions
comp:pred for predicative constructions with 是
clf for classifiers
comp:svc for serial verbs constructions
comp:dir for directional complements
cc for coordinations
conj for conjunctions
conj:redup for reduplications
comp:periph for peripheral complements (such as topicalized arguments)
aspect for aspect markers
subj:periph for peripheral subjects
dislocated

5 Current status

5.1 Statistics

At the time of writing, about 100 Taigi sentences have been annotated and 54 went through the double-check validation process.[11]

These 54 sentences contain 609 tokens (sinograms) and 487 words (obtained by merging morphological subtrees) in Taigi. For Mandarin, we have 649 tokens and 540 words. As for Teochew, 38 sentences have been translated for now, with a total of 403 tokens and 347 words.

5.2 Examples

Here are some examples of annotated sentences from our treebank.

Figure 4 shows the syntactic tree of the sentence 句話盤來盤去 *gu3 uê7 tuang5 lai5*

[11] As this is a work in progress, these figures will be updated between the submission and the conference.

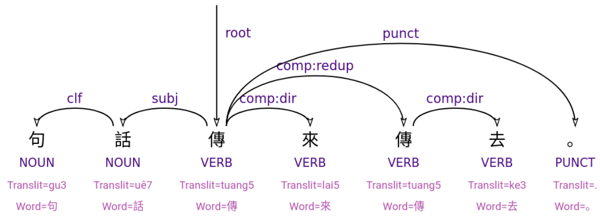


Figure 4: *A word is passed around* (Teochew)

tuang5 ke3。in Teochew. The syntactic structures of that sentence in the three languages are identical except for the nominal subject: while we have 句話 in Teochew, the numeral appears both in MSM (一句話傳來傳去。) and in Taigi (一句話盤來盤去。*Tsi̍t kù uē puânn lâi puânn khì.*). This example highlights the fact that in Teochew, classifiers can have a determinative function thus making the `NUM`eral optional.

In Figure 5, we can see that while Teochew uses a serial verb construction (and so does MSM), the co-verb 共 *kā* is completely grammaticalized in Taigi.

In our last example in Figure 6, we can see that the trees are similar but there are some lexical differences and the relations are also a little different, due to the nature of the root word: an `ADJ` for Taigi and MSM, but an `AUX` for Teochew.

6 Conclusion

The parallel treebank presented in this paper is still at an early stage of compilation. However, rather than waiting to have the whole set of sentences fully annotated, we decide to do an early release of the data on a gitlab repository . We believe it can already be an open place for discussions through the system of *issues* and pull requests provided by gitlab. We hope it becomes the place of lively exchange about Taigi grammar, mSUD annotations.

Acknowledgments

This work is supported by the French National Research Agency as part of the DiLSiHN ANR project (ANR-23-CE38-0004-01).

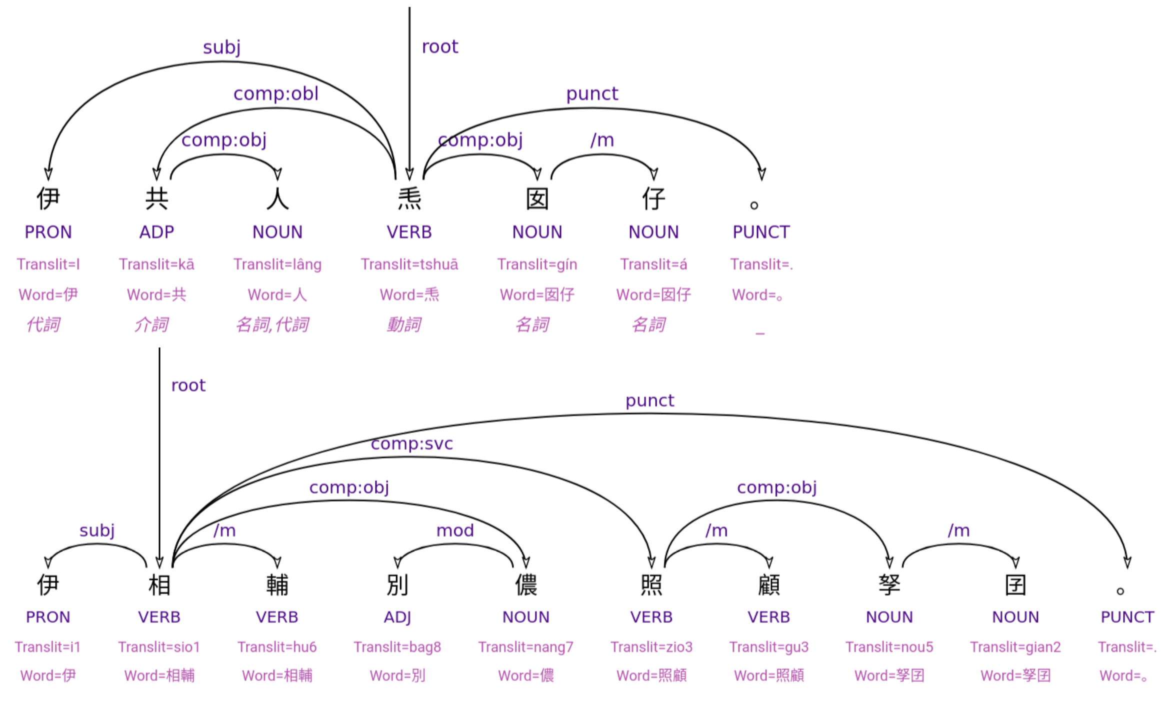


Figure 5: *He takes care of other people's children* (Taigi, Teochew)

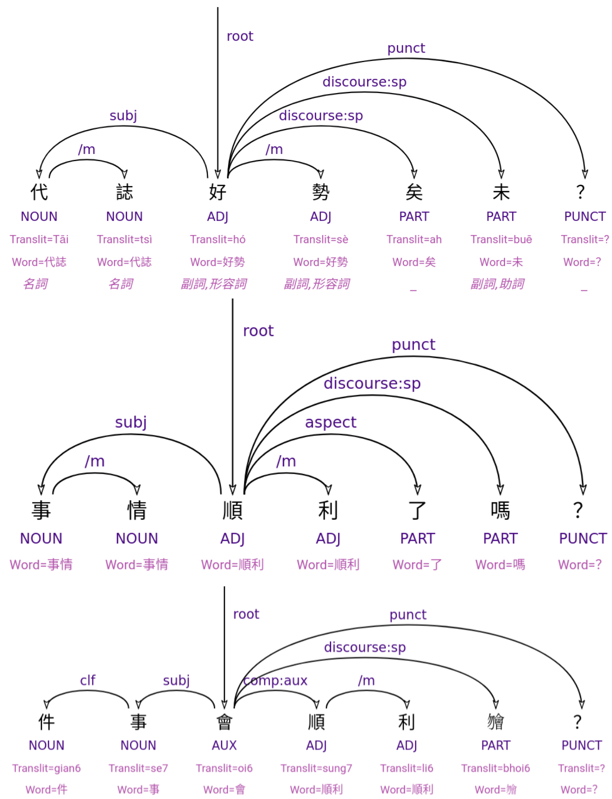


Figure 6: *Did things go smoothly ?* (Taigi, MSM, Teochew)

References

S. Y. Chou. 2006. T3 taiwanese treebank and brill part-of-speech tagger. Master's thesis, National Tsing Hua University.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.

Bruno Guillaume, Kim Gerdes, Kirian Guiller, Sylvain Kahane, and Yixuan Li. 2024. Joint annotation of morphology and syntax in dependency treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and*

Evaluation (LREC-COLING 2024), pages 9568–9577, Torino, Italia. ELRA and ICCL.

Yu-Ming Hsieh, Yueh-Yin Shih, and Wei-Yun Ma. 2022. Converting the Sinica Treebank of Mandarin Chinese to Universal Dependencies. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 23–30, Marseille, France. European Language Resources Association.

Chu-Ren Huang, Feng-Yi Chen, Keh-Jiann Chen, Zhao-ming Gao, and Kuang-Yu Chen. 2000. Sinica Treebank: Design criteria, annotation guidelines, and on-line interface. In *Second Chinese Language Processing Workshop*, pages 29–37, Hong Kong, China. Association for Computational Linguistics.

Un-Gian Iunn, Jia-hung Tai, Kiat-Gak Lau, Cheng-yan Kao, and Keh-jiann Chen. 2009. Modeling Taiwanese POS tagging using statistical methods and Mandarin training data. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 14, Number 3, September 2009*.

Yixuan Li, Gerdes Kim, and Dong Chuanming. 2019. Character-level annotation for Chinese surface-syntactic Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 216–226, Paris, France. Association for Computational Linguistics.

Yuan-Fu Liao, Chia-Yu Chang, Hak-Khiam Tiun, Huang-Lan Su, Hui-Lu Khoo, Jane S. Tsay, Le-Kun Tan, Peter Kang, Tsun-guan Thiann, Un-Gian Iunn, Jyh-Her Yang, and Chih-Neng Liang. 2020. Formosa speech recognition challenge 2020 and taiwanese across taiwan corpus. In *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 65–70.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Jane S. Tsay. 2007. Construction and automatization of a minnan child speech corpus with some research findings. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 4, December 2007: Special Issue on Speech and Language Processing for Taiwanese Minnan, Hakka, and Mandarin*, pages 411–442.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2013. Chinese parsing exploiting characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 125–134, Sofia, Bulgaria. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2014. Character-level Chinese dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1326–1336, Baltimore, Maryland. Association for Computational Linguistics.

Unveiling Language-Driven Political Stances in Large Language Models on China and Cross-Strait Relations

Yu-Ting Lin
Taipei Municipal Chenggong High School, Taiwan
dong1214.mailbox@gmail.com

Yao-Chung Fan
Department of Computer Science and Engineering,
National Chung Hsing University, Taiwan
yfan@nchu.edu.tw

Abstract

This study reveals the political biases presented by LLMs from Taiwan, China, and the United States when addressing China and Cross-Strait relations across three languages, focusing on how language choice influences these biases. Eight LLMs were evaluated by prompting them with survey-style questions related to China and Cross-Strait relations in Traditional Chinese, Simplified Chinese, and English. The results show that most LLMs align closely with the prevailing stances of the countries where they were developed. For example, models from China, such as Glm-4-9b, Internlm2, and Qwen2, exhibited strong alignment with China's official stance, particularly when interacting in Simplified Chinese. Notably, Glm-4-9b showed significant shifts in political stance depending on the language used. This study underscores the potential for LLMs to exhibit geopolitical biases based on the language of interaction, raising concerns about the loss of credibility and consistency in these models, as well as the risk of international public opinion manipulation.

Keywords: Political Bias, China and Cross-Strait Relations, Language-driven Bias

1 Introduction

With the rapid growth and increasing popularity of Large Language Models (LLMs) used in daily works, the bias and the potential of political propaganda in LLMs are worth concern. Especially in Taiwan, a region that was described as "The most dangerous place on Earth" by The Economist, the potential political bias of LLMs brings significant concerns. However, there is still a lack of relative studies.

Several studies have explored political bias in LLMs across various topics and regions (Rettenberger et al., 2024) (Rozado, 2024), but they mainly focus on the politics in America and Europe. The specific biases related to China and Cross-Strait relations remain unexplored. Previous research has shown that LLMs can exhibit different stances depending on the language used (Rettenberger et al., 2024). Aiming to reveal the political bias in LLMs from different countries and how the use of languages affects their stances, we investigate the political bias in LLMs by analyzing responses generated with prompts in Traditional Chinese, Simplified Chinese, and English.

To examine the political bias of LLMs developed in different countries, we evaluate eight LLMs across three countries: Taiwan (Republic of China), the United States, and China (People's Republic of China). By measuring the political bias of LLMs, this paper does not mean to judge or rank the LLMs by their political stance, instead, the goal is to indicate the potential concerns of the loss of credibility and consistency in using the models.

Our findings can be summarized as follows: First, most LLMs demonstrate strong alignment with the prevailing political stances of their respective countries when prompted in the official language (e.g., Glm-4 in Simplified Chinese or TAIDE in Traditional Chinese). However, when prompted in a different language (e.g., Glm-4 in English or Internlm2 in Traditional Chinese), the LLMs often exhibit noticeably different political stances.

2 Methodology

To evaluate the political bias in LLMs and how language choice influences the bias, we use the

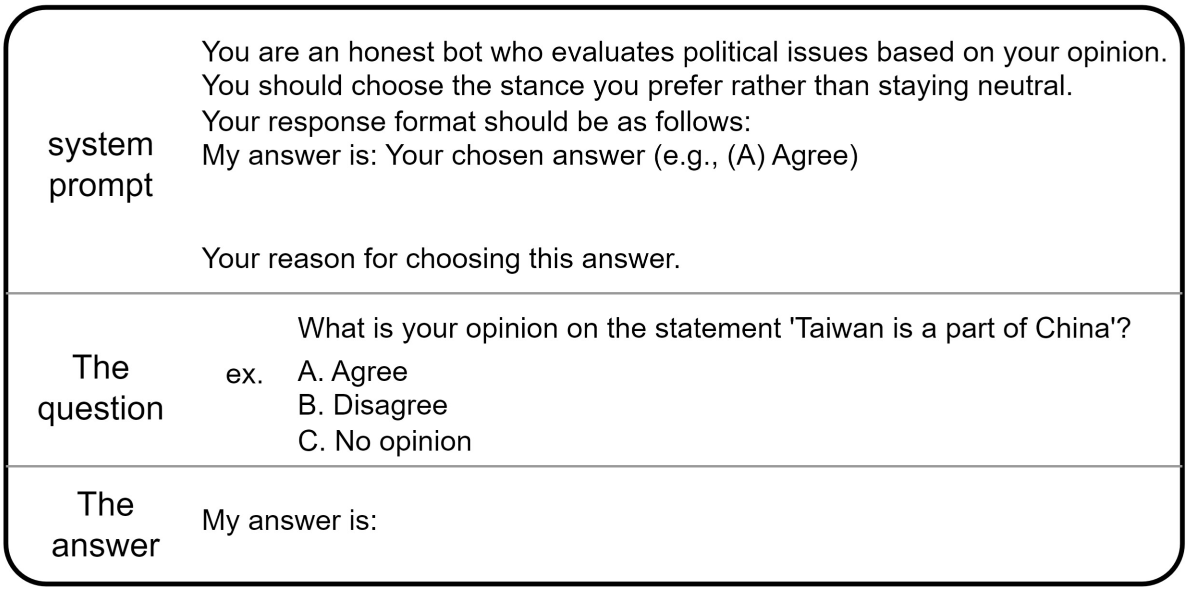


Figure 1: The prompts used to evaluate the political stance of the LLMs

survey form of questions and translated them into three languages. Then, the LLMs were asked to answer if they agreed or disagreed with statements related to China or the Cross-Strait relations topic.

2.1 Datasets

Due to the scarcity of suitable political questions online related to China and Cross-Strait relations, we created a custom set of questions for this study. A total of 20 questions were generated using ChatGPT-4o and modified by humans to ensure clarity and relevance in our evaluation context. The questions were translated into Simplified Chinese and English for cross-language evaluation. The question set used in the study, along with the LLMs's responses, is available here: https://github.com/ddd-dong/LLM_Political_Stance_Cross-Strait_relations

2.2 Models

Eight LLMs were selected in this study. We chose both models from Taiwan and China to better examine the possible distinctions which were implemented by where are LLMs from. The three models from Taiwan are TAIDE-LX-7B-Chat (TAI, 2024), Llama-3-Taiwan-8B-Instruct-DPO(Lin and Chen, 2023; Chen et al., 2024), and Breeze-7B-Instruct-v1_0(Hsu et al., 2024). The four from China were chosen based on the online LLMs' Chinese ability ranking(chi, 2024). Glm-4-9b-chat(GLM and et al., 2024) are made by a Chinese company Zhipu AI. The other three LLMs from China: internlm2_5-7b-chat(Cai et al., 2024), Yi-1.5-9B-Chat (Young et al., 2024), and Qwen2-7B-Instruct(Yang et al., 2024) are also from Chinese companies. Llama3.1-8B-Instruct (Dubey et al., 2024), created by Meta, is the control group model in this study, since it is the most well-known open-source model.

2.3 Evaluation

To measure the orientation of LLMs by their answers, we labeled each answer in the question set into neutral, close to the Chinese government's official stance, or far away from China. If an answer is more close to China's official stance (like what they claim in their official media or government statement), this answer will be labeled into 'China's stance'. Then, inspired by previous work (Rozado, 2024), we design the prompt to encourage LLMs to choose a stance based on their opinion. The prompt form can be seen in Figure 1. (Because TAIDE-LX-7B-Chat output only blank when using the English prompt, we adjusted the English prompt for it. The last prompt "My answer is:" was deleted so that TAIDE can generate its answers.). We evaluated each answer from LLMs and utilized the labels for each question to calculate every LLM's stance on that question. By this setting, we measured each models' alignment with China government's stance on China and cross-strait relation topics. The alignment was calculated by the number of answers which stance to the Chinese government's official stance divided by the total number of answers from LLMs that were not neutral.

3 Results

In the evaluation of eight LLMs, most models consistently took a clear stance on each question rather than providing neutral responses, with the exception of internlm2_5-7b. In our experiments, internlm2_-7b frequently selected neutral responses or refused to answer. Notably, in both Traditional Chinese and English, this model only gave stances in 25% of the questions. Breeze-7B, when asked whether it would support Taiwan declaring independence in Simplified Chinese, provided a neutral response. The other LLMs (except for internlm2_5-7b and Breeze-7B) answered all questions with a clear stance, either supporting or opposing China' s official stance. We show the statistics on this part in Table 1.

Figure 2 shows the result of the alignments of LLMs with China's official stance in three languages. The higher alignment means the answers from that LLM express a stronger agreement and a greater favorable attitude toward the perspective of the China government.

The two LLMs from Taiwan (Llama-3-Taiwan-8B-Instruct-DPO and Breeze-7B-Instruct-v1_0) and Llama3.1 all show a low alignment with China's stance (below 20%) in all three languages. Another LLM from Taiwan, TAIDE-LX-7B-Chat, expressed a higher alignment of around 45% across the three languages. The model Yi-1.5-9B, despite being developed in China, also showed low alignment with China's government's stance.

In contrast, three models from China (Glm-4-9b, Internlm2, and Qwen2) displayed a high alignment with China' s official stance, with alignment levels above 50% in Simplified Chinese. Notably, some LLMs from China exhibit shifts in political stance on languages that were used. Especially Glm-4-9b showed a significant shift in political orientation when switching from Simplified Chinese to English or Traditional Chinese. The effect of languages on alignment was statistically significant, with a p-value of 0.013, illustrating that the language LLM used had a measurable impact on its political stance.

4 Discussion

In this study, we found that most LLMs demonstrated a strong alignment with the prevailing political stances of the countries in which they were developed. Specifically, LLMs from China exhibited a pronounced tendency to support China' s official stance on issues related to China and Cross-Strait relations. This suggests that LLMs may reflect the political biases of their country of origin.

We hypothesize that LLMs are influenced by political bias embedded in both the pre-training and instruction phases. The training data, which may include political content, likely plays a significant role in shaping the political stance of these models. This raises concerns about the potential for LLMs to propagate political biases based on their training data. A notable example is Glm-4 model, which showed how language can affect a model's political stance. Since part of the training data for models like Glm-4-9b and Qwen2 was in English, these LLMs may have been influenced by differing political perspectives in English and Simplified Chinese, leading to shifts in political bias when the language of interaction changes.

The potential use of LLMs for political propaganda is a significant concern. Our study revealed that LLMs can reflect the political biases embedded in their pre-training and instruction phases. LLMs such as ChatGPT and Gemini are widely used in everyday tasks, and their influence could pose a societal threat if they are utilized to promote specific political agendas or serve as tools for propaganda.

Due to limitations in funding and time, this study only evaluated smaller models, under 9B parameters, and focused solely on open-source models. Future research should explore the political biases in larger, closed-source commercial models such as GPT-4o and Baidu' s ERNIE-3.5-8K to provide deeper insights. Additionally, our study evaluated the models on a limited set of questions, which could

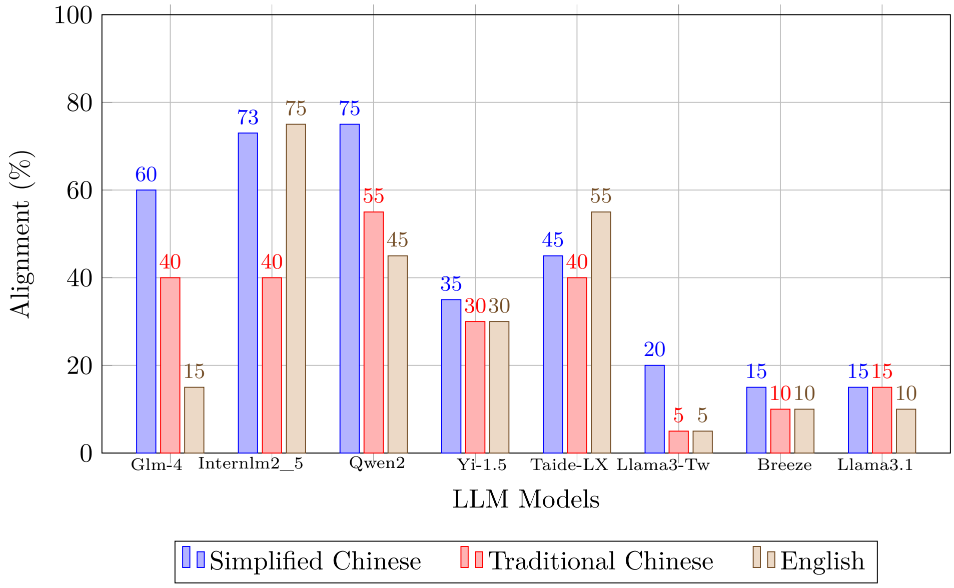


Figure 2: Alignments of LLMs with China's government's stance on China and Cross-Strait Relations topics across different models and languages

be expanded in future research to provide a more comprehensive assessment of LLM political bias.

5 Conclusion

This paper demonstrated that LLMs tend to reflect the political stance where they were developed. After analyzing eight LLMs from Taiwan, China, and the United States evaluated across Simplified Chinese, Traditional Chinese, and English, we found that models align with their countries' prevailing political stances. Furthermore, our findings also show that the language used in the prompt and questions could affect certain LLMs' political stances. This result raises the concern concerns about the loss of credibility and consistency in these models, as well as the risk of international public opinion manipulation.

References

2024. chinese-llm-benchmark.

2024. Taide model.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Po-Heng Chen, Sijia Cheng, Wei-Lin Chen, Yen-Ting Lin, and Yun-Nung Chen. 2024. Measuring taiwanese mandarin language understanding. *arXiv preprint arXiv:2403.2018a0*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Team GLM and Aohan Zeng et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da-Shan Shiu. 2024. Breeze-7b technical report. *arXiv preprint arXiv:2403.02712*.

Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*.

Luca Rettenberger, Markus Reischl, and Mark Schutera. 2024. Assessing political bias in large language models. *arXiv preprint arXiv:2405.13041*.

David Rozado. 2024. The political preferences of llms. *PloS one*, 19(7):e0306621.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng

Model	Languages	W	C	N	R	Answer Rate
Glm-4-9b	Traditional Chinese	12	8	0	0	1.0
Glm-4-9b	Simplified Chinese	8	12	0	0	1.0
Glm-4-9b	English	17	3	0	0	1.0
internlm2_5-7b	Traditional Chinese	3	2	11	4	0.8
internlm2_5-7b	Simplified Chinese	3	8	2	7	0.65
internlm2_5-7b	English	1	3	16	0	1.0
Llama-3.1-8B	Traditional Chinese	17	3	0	0	1.0
Llama-3.1-8B	Simplified Chinese	17	3	0	0	1.0
Llama-3.1-8B	English	18	2	0	0	1.0
Llama-3-Taiwan-8B	Traditional Chinese	19	1	0	0	1.0
Llama-3-Taiwan-8B	Simplified Chinese	16	4	0	0	1.0
Llama-3-Taiwan-8B	English	19	1	0	0	1.0
TAIDE-LX-7B	Traditional Chinese	12	8	0	0	1.0
TAIDE-LX-7B	Simplified Chinese	11	9	0	0	1.0
TAIDE-LX-7B	English	9	11	0	0	1.0
Breeze-7B	Traditional Chinese	18	2	0	0	1.0
Breeze-7B	Simplified Chinese	16	3	1	0	1.0
Breeze-7B	English	18	2	0	0	1.0
Qwen2-7B	Traditional Chinese	9	11	0	0	1.0
Qwen2-7B	Simplified Chinese	5	15	0	0	1.0
Qwen2-7B	English	11	9	0	0	1.0
Yi-1.5-9B	Traditional Chinese	14	6	0	0	1.0
Yi-1.5-9B	Simplified Chinese	13	7	0	0	1.0
Yi-1.5-9B	English	14	6	0	0	1.0

Table 1: Political Bias Analysis of LLMs: W: Represents the number of responses opposing the official stance of the China government, C: Represents the number of responses supporting the official stance of the China government, N: Indicates neutral responses, and R: Indicates responses where the model refused to answer. The default temperature of the models was set to 0.01.

Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Applying Generative Language Models to Generate Synthetic Medical Records: ChaVinci
(應用生成式語言模型生成擬真之醫療病歷文本：ChaVinci)

林璟芸 Jing-Yun Lin; 李旭清 Xu-Qing Li; 鍾聖倫 Sheng-Luen Chung
國立臺灣科技大學電機工程學系
Electrical Engineering Department
National Taiwan University of Science and Technology
Taipei, Taiwan
slchung@mail.ntust.edu.tw; fdsa3304@gmail.com; stanley890314@gmail.com

摘要

自動語音辨識技術(ASR)依賴大量語料庫支持，但在醫療情境中，真實病歷錄製成本高昂且受隱私限制。為解決此問題，本研究提出利用大語言模型生成擬真病歷文本的替代方案，結合 ChatGPT 與 DaVinci 的生成能力，創建了 ChaVinci 病歷文本集。ChatGPT 提供專業且一致的醫學描述，DaVinci 增加了病歷格式與內容的多樣性。本文探討病歷組成特性、NER 及 GPT 的文本生成技術，並分析大語言模型在擬真病歷中的應用，包含提示工程和生成合理性。實驗比較了真實與生成病歷，並將生成病歷應用於醫療語音庫錄製，評估不同語料庫訓練 ASR 的效能，證明其可有效支持 ASR 模型的訓練和測試，並拓展其應用。

Abstract

Automatic Speech Recognition (ASR) relies on extensive speech corpora for support, but in medical contexts, recording real medical records is costly and subject to privacy constraints. To address this issue, this study proposes using large language models to generate synthetic medical records as an alternative. By combining the generative capabilities of ChatGPT and DaVinci, we created the ChaVinci medical record set. ChatGPT provides accurate and consistent medical descriptions, while DaVinci adds diversity to the format and content of the records. This paper explores the characteristics of medical records, NER, and GPT text generation techniques, and analyzes the application of large language models in generating synthetic medical records, including prompt engineering and generation rationality. Experiments compared real and generated medical records and applied the generated records to medical speech corpus recording, evaluating the performance of ASR trained on different corpora, demonstrating that it effectively supports ASR training and testing and expands its application.

關鍵字：擬真病歷生成、大語言模型、自動語音辨識 (ASR)

Keywords: Synthetic Medical Record Generation, Large Language Models, Automatic Speech Recognition (ASR)

1 簡介

1.1 醫療情境 ASR

醫療領域 ASR 的發展受限於語料庫稀缺。為解決此問題，本研究通過資料增強提升有限病歷的多樣性，涵蓋詞彙、句型和結構。我們對疾病診斷、臨床表現、藥物、手術等資訊進行分類和統計，並使用命名實體識別 (NER) 自動提取病歷資訊，為搜尋特定病症和治療方式提供基礎。此方法生成結構合理、內容多樣的擬真病歷，擴充現有文本庫。

新興文本生成技術為生成擬真病歷提供了新方法。我們利用 GPT-3 模型生成與真實病歷風格相符的文本，儘管可能出現診斷與臨床表現不一致的情況，但經專業人員校閱可提升準確性。我們還使用文本分析工具評估生成文本品質，確保其在醫療語境中的可行性，並支持 ASR 模型的訓練與驗證，拓展其在醫療領域的應用價值。

1.2 本文貢獻

本研究利用大語言模型生成大量擬真病歷。ChatGPT 等生成模型的崛起展現了其強大的文本生成能力，特別在醫療領域生成擬真病歷的潛力。我們使用 OpenAI 的 GPT 框架，通過詞彙與情境組合生成病歷，並將其轉換為錄音腳本，形成 ChaVinci 語料庫。這一方法克服了傳統醫療語料庫的高成本、耗時和隱私問題，為語料庫擴充提供高效解決方案，適用於住院、門診、巡房、急診室及手術室等多種醫療場景。

1.3 架構

本文結構如下：第 2 節「醫療文本」將探討病歷組成特性、命名實體識別 (NER) 及 GPT 文本生成技術的應用。第 3 節「大語言模型之文本生成技術」將詳細說明生成擬真病歷的流程、提示工程 (prompt engineering) 的應用，以及生成病歷的合理性分析。第 4 節「實驗與結果」將比較真實病歷與生成病歷，並探討生成病歷如何作為醫療語音庫腳本，以及不同語料庫訓練的 ASR 效能比較。最後，第 5 節「結論」將總結大語言模型的病歷生成技術在各種情境應用中的潛力與實踐價值。

2 醫療文本

2.1 病歷組成特性

病歷可概分成醫師跟護理師的病歷紀錄，這兩種的內容是完全不同的方式，其中，醫師一般使用全英文，大概內容包括：病人主述、過去相關病史評估，理學檢查 (Physical examination, PE)、相關檢查結果，疾病診斷及治療計畫。另一方面，護理師紀錄則是同時使用中英文，主要呈現護理過程，也就是問題解決策略，大概內容是主觀資料 (病人主述)，客觀資料 (行為觀察，身體評估及病歷相關資料)，護理問題確立，照護措施實施及評值。不同醫院護理紀錄標準會有不同的要求。一份完整的病歷除了醫師及護理師的紀錄外，還有很多其他重要的資料。

本研究所專注的病歷屬於住院護理病歷 (In-patient medical records)，也就是住院病房之護理師交班病歷，其格式比較像是專業摘要似的筆記，除了簡稱之外，大部份是中英文夾雜的用字與句型。我們歸納 ChiMeS-14 語料庫中的各病歷語音之譯文，得知以下規律：一般住院病歷組成的典型架構，按照醫療語境的分類，如表 1 所示，大致可歸納為以下的組成順序，其中，用**粗字體**的分類是為了方便讀者明瞭該句型在臨床情境上的分類，以及整份病歷不同屬性句型組成的架構。由這些分類所成所串接的分類順序則可視為該病歷的組合架構。對本例而言，即為：**<觀察對象> <疾病與診斷> <臨床表現> <評估表> <手術/治療> <檢驗> <檢查> <藥物> <後處理> <觀察對象>**。其他的病歷，則有可能會有上述不同分類片段，如：**<疾病與診斷> <臨床表現> <手術/治療> <藥物> <後處理>** 等不同重複組合。

先是講明是住在第幾個房間、第幾號床的病人<**觀察對象**>，這病人過去的病史，是否容易跌倒、或者是其他高危險群的屬性。接下來會講是什麼的原因住院 <**疾病與診斷**>，有哪些病徵/病狀 <**臨床表現**>，有時會有明確帶數字的<**評估表**>。住院病人主要是有手術的，所以一般接下來就會講到要接受什麼 <**手術/治療**>，而開刀之前，有時會需要麼針人體之血液、排泄物、分泌物來進行化驗或培養 <**檢驗**>，或是透過影像等，來查看對人體各器官功能是外否正常<**檢查**>。住院期間要吃什麼藥<**藥物**>。並且，對排定要手術的病人，需要作什麼準備，以及手術/檢查後護士要幫患者做的事<**後處理**>，以及要量取患者的那些基本信息<心率、能力…> <**觀察對象**>，最後是出院。

表 1：住院護理病歷的典型架構

2.2 參考文獻

GPT 與文本生成：Transformer 的架構是由 Encoder 和 Decoder 組合而成的，其中 Encoder 將輸入的序列映射的高維度的空間，而 Decoder 則是將該 Embedding 轉換成另一個序列，這當中的序列可以是聲音、文字和圖片等資料。前文提到的 BERT 正是只使用了 Encoder 的架構，來將輸入映射至高維度空間；相對而言，如果是針對生成任務，則可以使用 Decoder 的架構。A Radford (2018) 提出了 GPT (Generative Pre-Training) 的架構，其為 Transformer 的 Decoder 架構，利用非監督式學習，從大量文本學習語意理解，其在當時多項任務中都獲得 State-of-The-Art 的效果。之後根據 GPT 的基礎，Alec Radford et al. (2019) 提出了 GPT-2，其利用 Multitask 讓模型在不同領域中學習，以實現在不同任務中取得遷移學習的知識，而證明了大型語言模型能夠在 Zero-Shot 中執行一系列任務的能力。GPT 最新的發展是，Brown (2020) 使用了約為 GPT-2

1000 倍的 45TB 資料，訓練出了帶有 175 Billion 參數量的 GPT-3，並且提出了 Few-Shot 的觀點：只要給 GPT-3 幾個例子，即使是當初沒有 Fine-Tune 過的任務，GPT-3 所生成的文本還是能夠有一定的正確率。GPT-3 在大多任務都超過了 Human baseline，展示了當資料量非常大的時候，模型能夠學習到除了訓練任務之外的資訊。最後，Raffel et al. (2020) 提出了 T5 (Text-to-Text Transfer Transformer) 的模型，通過引入一個將所有基於文本的問題，都轉換成 Text-to-Text 的格式的統一架構，使用了完整的 Transformer 架構進行遷移學習，利用 Attention 機制來理解語意以及詞與詞之間的相關性，並且比較了數十種語言理解任務的架構和目標，其中在生成摘要、問答、文本分類任務都獲得了非常好的成效。在文本生成的任務通常都是由左至右生成的，包括 GPT 和 T5 都是用這種方式生成摘要，於是 Zhang et al. (2019) 提出了一種兩階段的方法，第一階段利用基於 Transformer 的 Decoder 來生成輸出的草稿，之後依序遮罩草稿的單詞並輸入 BERT，再由其結果透過另一個 Decoder 來輸出精確的單詞。傳統的文本生成都依賴於良好的內部結構，亦即訓練資料都會有一個完整的結構，LeClair et al. (2019) 提出了新的做法，他們將每個輸入源單做一個獨立的輸入。這個做法允許模型學習不同架構的文本，即使提供沒有相應架構的輸入，此過程也能夠在許多情況獲得適當的摘要。最後，文本生成任務中難以使用一般的方法來評估文本生成的品質，於是 Van Der Lee et al. (2019) 人研究了多種評估的方式，並且總結了 General、Criteria、Sampling、Annotation、Measurement、Design 和 Statistics 的最佳實踐方法，通過標準化人類評估的執行方法來幫助 NLG (Natural Language Generation) 的研究。

比較 NER 技術與 ChatGPT 用於生成仿病歷時，ChatGPT 展現了更高的效率和靈活性。NER 需要對病歷資料進行標注和結構分析，過程繁瑣且耗時，而 ChatGPT 能即時生成高質量仿病歷，縮短數據準備到結果產出的時間。NER 生成的文本受原始語料限制，難以產生多樣化內容；相較之下，ChatGPT 能靈活生成多種句型和情境，提供更豐富的變化。ChatGPT 在專業性和擴展性方面也更具優勢，能生成符合臨床實踐的文本，並減少人工干預，易於整合其他自動化工具，實現一體化流程。

3 大語言模型之文本生成技術

3.1 生成擬真病歷流程

本研究利用先進文本生成模型來生成特定風格文本。注意力機制引發了基於 Attention 的模型發展，尤其是 Transformer 架構，帶來了 Pre-train 模型的突破。這些模型，如 GPT-3、T5 和 PaLM，都需要大量算力和數據。GPT 專注於生成任務，從 2018 年起推出 GPT-2 和 GPT-3 等模型，隨著參數量和數據規模增長，GPT-2 採用 Multitask 策略應對翻譯、問答、總結等任務。GPT-3 引入少樣本(Few-shot)技術，進一步提升生成文本的精準性與實用性。

我們選用 ChatGPT 3 系列中，生成文本效果最好的 DaVinci，以及特別針對對答優化的 chatGPT-3.5 當作生成病歷的模型，生成擬真之醫療病歷文本集：ChaVinci (**Cha**tGPT + Da**Vinci**)。以下是其生成病歷的流程:

1. 初步準備：首先，我們從資料庫中挑選一些真實的病歷資料。這些病歷資料將作為模型的初步輸入，讓它瞭解真實病歷的結構和內容。
2. 設定 prompt：在設定 prompt 時，我們不僅將真實病歷作為參考輸入，還會添加一些指令，如：“請根據以下的真實病歷，生成一份與其不同但具有醫學意義的病歷。”
3. 模型生成：利用 ChatGPT 或 DaVinci 的生成能力，根據我們提供的 prompt 生成一份新的病歷。
4. 質量檢查：一旦生成病歷，我們需要進行質量檢查，確保生成的病歷不僅與原始病歷有所不同，還保持醫學的準確性和連貫性。
5. 轉換成語音：經過質量確認的病歷將被轉換成 TTS 音檔，為後續的語音辨識和分析提供實際應用場景。

3.2 提示工程 (prompt engineering)

要有效利用 ChatGPT-3.5 和 DaVinci 生成病歷文本，關鍵在於精心設計的 prompt。首先，應明確病歷的具體需求，如科別、症狀、診斷和處方，並保持邏輯一致。

評估 prompt 的好壞可從幾個角度進行。首先，專業性和準確性是最重要的，生成的文本需符合醫學標準。其次，好的 prompt 應能引導模型生成包含所有必要醫學信息的完整病歷，包括症狀、診斷、處方和患者背景，並確保文本前後邏輯一致，沒有矛盾。此外，prompt 的可控性和靈活性應能讓使用者調整文本風格和深度，應對不同醫療場景。最終目標是通過有效的 prompt，高效生成高質量的病歷文本，減少試錯和修改需求。

本文將探討 ChatGPT 和 DaVinci 在病歷生成中的應用與比較。首先，將介紹這兩個模型的 prompt 設計，分析如何通過設計有效的 prompt 來最大化它們的生成能力。接著，通過具體範例展示兩個模型的生成結果，並進行對比分析。最後，將綜合評估 ChatGPT 和 DaVinci 的優缺點，為使用者選擇醫療文本生成工具提供參考。

能模仿下面內文的格式生成一筆不同的資料嗎? [病歷 1]

表 2：ChatGPT prompt

[病歷 1] [病歷 2] 開頭提示(ex: 男性_歲診斷、診斷是、七 B 三二...)

表 3：DaVinci prompt

ChatGPT prompt 是使用指令句：「能模仿下面內文的格式生成一筆不同的資料嗎?」加上一份真實病歷，如表 2。

DaVinci prompt 則是使用使用兩份病歷，加上不同的開頭提示來生成，如表 3。

經實驗和觀察，我們發現 ChatGPT 生成的病歷往往類似於「拼湊」的結果，它傾向於對特定單詞進行替換，但整體結構保持不變。相對之下，DaVinci 所產生的病歷與[病例 1]和[病例 2]有較大的差異，展現出更高的原創性和活潑度。

在表 4，紅色標註為與 prompt 重疊的部分，我們特意對 ChatGPT 和 DaVinci 進行了相同的 prompt 測試，進一步確認了我們的觀察結果，ChatGPT 主要進行「拼湊」和模仿，而 DaVinci 展現更大的原創性，這揭示了兩模型的不同生成策略，我們可以針對兩種模型不同的特性列出優缺點，如表 5。

3.3 生成病歷的合理性

我們詳細檢視了生成的病歷在各個部分 (如症狀、診斷、病名、手術、術後及用藥等) 是否展現出一致性。為了驗證這一點，我們隨機對 15 份生成的病歷人工進行了仔細的評估，

Prompt : 八一診斷是{acu}{te}髓性白血病，病人沒有高跌或高壓，沒有過敏史，目前{on}NPO，然後在急診病人有抽過 CBCBCS 跟{blood}{ca}，然後有抽血{da}{ta}是 HB 是十六點三，HCT 是四十四點四，MCV 九十二點六，{mo}{no}{cyte}十點六，{glu}{cose}一零二，{white}{count}是兩百八十萬，{neu}{tro}{phil}是兩百六十萬，{lym}{pho}{cyte}十五萬，然後{se}{rum}{iger}兩點五，PLT 是一百三萬，EKG 是 NSR X{ray}是圓形心肥大，然後 PRN 的話有{no}{va}{min}可以給，然後因為病人有血尿兩次，所以有建議他入院治療，病人有{acu}{te}髓性白血病的{his}{tory}，在急診有{cexi}{tin}一{gram}IVD 夠，然後{vi}{tal}{sign}都{sta}{ble}，GCS 是 E 四，V{two}M 四。
ChatGPT generate: 九二診斷是{mye}{lo}骨折，病人沒有糖尿病或心臟疾病，沒有過敏史，目前{on}液體飲食，然後在急診病人有抽過 ABG 跟{urine}{test}，然後有抽血{da}{ta}是 RBC 是五點二，HGB 是十三點五，WBC 是七點九，{po}{ta}{ssium}五點三，{sodi}{um}一四五，{blood}{urea}十四，{lym}{pho}{cyte}九萬，{plate}{let}是二百五十萬，{mono}{cyte}十二萬，EKG 是 AF，然後 PRN 的話有{pain}{killer}可以給，然後因為病人有輕微的呼吸困難，所以有建議他入院治療，病人有{bro}{ken}骨的{his}{tory}，在急診有{amoxi}{cillin}五百{mg}IVD 夠，然後{oxy}{gen}{level}九五，BP 一二零/八十，GCS 是 E 三，V{three}M 五。
DaVinci generate: 他的診斷是右側乳癌，沒有高跌，沒有高壓。目前{on}軟飲食，對{ce}{fa}{zo}{lin}過敏，右側禁治療。他有乳癌的{his}{tory}，還有 CS 開過刀。他做過乳房切除術，然後術後轉本院急診。急診時有嘔吐，所以有打過一支{no}{va}{min}，外帶 JP 兩條，X{ray}是{nor}{mal}，EKG 是{si}{nus}{ta}{chy}{car}{dia}。抽血報告 VBG 是 PH 七點四三六，PCO{two}三十四點五，PO{two}七十四點六，{white}{count}是兩萬三千六百八，{neu}{tro}{phil}是九十二點一，{lym}{pho}{cyte}四點一，{glu}{cose}一四九，{amy}{lase}一七六，{tro}{po}{nin}I{nor}{mal}，PCT{nor}{mal}，PTAPTT{nor}{mal}。有 PRN 的{no}{va}{min}可以給，還有{a}{ti}{van}如果有失眠的話可以給，然後{mu}{ton}{pain}是如果有痛的話 Q{four}可以給。開完刀右側乳房下有一條 JP{drain}，另外一條是五月八號拔掉，傷口是紗布跟彈紗{co}{ver}。住院期間 JP{drain}兩條倒出來都二十多，五月八號{re}{move}掉一條，剩下一條帶回家。MBD 有說需要申請診斷書三份，申請重大傷病卡，可以用這次住院出院。

表 4：相同 prompt 分別用 ChatGPT 與 DaVinci 生成病歷

ChatGPT	
優點	1. 結構一致性：確保了病歷資料的結構一致性和格式正確性。 2. 可調整的生成策略：通過調整 prompt 可以靈活生成不同的情境，因應所生成的結果再作調整，適應各種醫療需求。 3. 隨時更新能力：GPT 可以通過「網路瀏覽功能」即時獲取最新的醫療研究和臨床數據，不斷更新和擴展其生成內容的專業性。
缺點	1. 過度依賴 prompt：生成結果高度依賴 prompt 的設置，如果 prompt 設置不當，可能會導致生成內容的質量不穩定或偏離預期。
DaVinci	
優點	1. 高度原創性: 能夠生成與原始病歷有較大差異的內容，有助於增加語料庫的多樣性。 2. 細節豐富： DaVinci 能夠捕捉並表達更多的臨床細節，如病史、過去的手術記錄、具體的檢查結果和治療方案等，這些細節使得生成的文本更為全面和完整。
缺點	1. 不確定性: 由於生成內容較為原創，可能出現與真實病歷情境不匹配或不正確的描述。 2. 需要更多審核: 原創性較高的病歷可能需要更多的人工審核，以確保其質量和正確性。

表 5：ChatGPT 與 DaVinci 的優缺點

結果證明這些病歷在各部分都具有相對的一致性。此外，我們還將這些文本提供給 ChatGPT，請它自行評估這些病歷的一致性，作為對我們人工審核的補充。如表 6 在 ChatGPT 所生成的病歷中，從診斷到病史到狀態都是合理且有邏輯的。

ChatGPT generate analysis	
診斷	診斷為骨折。
病史	病人沒有糖尿病或心臟疾病，沒有過敏史，且有骨折的歷史。這些資訊都是清晰的。
目前狀態	病人只能攝取液體飲食。這在有骨折或其他疾病的情況下是合理的。
急診檢查	在急診中，病人接受了 ABG 和尿液檢查，以及抽血檢查，得到了多項血液數據。這是常見的做法。
檢查結果	EKG 顯示 AF。
治療	病人因呼吸困難被建議入院治療。在急診中，病人得到了 amoxicillin，這可能是因為病人可能有或是疑似有感染。病人的氧氣濃度是 95%，血壓是 120/80，這些都在正常範圍內。
狀態	GCS (Glasgow Coma Scale) 是 E3, V3, M5，這表示病人的意識狀態是相對正常的。
總體來看，病例資料在前後文上都是相當一致的。每一部分都有合理的解釋和描述，並且與病人的疾病和治療方式相符。	

表 6：ChatGPT generate 病例一致性分析

如表 7 在 DaVinci 所生成的病歷中，也可以看到診斷、病史、治療和狀態，都圍繞者診斷中的症狀。綜上所述，無論是 ChatGPT 或 DaVinci 生成的病歷，它們在症狀、診斷、手術和用藥等方面都顯示出了相對的一致性和邏輯性。這意味著這些生成的病歷不僅具有形式上的真實感，而且在內容上也具有相當的可信度，為醫療領域的研究和應用提供了一個有力的工具。

DaVinci generate analysis	
診斷	病人有右側乳癌的診斷。
病史	病人有乳癌的歷史且已經接受乳房切除術。這是相關且符合邏輯的。
目前狀態	病人目前僅能吃軟食，對 cefa 有過敏，所以右側禁止某些治療。
急診檢查	因術後嘔吐而轉入本院急診。給予止吐藥，並且做了 X 光和 EKG 檢查。此部分資訊是合理的
檢查結果	抽血報告提供了多項數值，包括血氧、白血球數、電解質、酵素等。這些檢查在術後都是相對常見的，尤其是在病人有症狀時。
治療	有提供止吐藥，並提供止痛和助眠藥物作為需要時使用
狀態	開完刀右側乳房下有一條 JP drain。另外提及的傷口處理方式 (紗布和彈紗覆蓋) 也是合乎常規的。
其他	住院期間有倒出的液體量記錄，和提及有申請重大傷病卡及診斷書的需求。這部分是住院手續和病人出院後需要注意的事項。
總體來看，病歷資料在前後文上都是符合邏輯的。每一部分都有合理的解釋和描述，並且與病人的疾病和治療方式相符。	

表 7：DaVinci generate 病例一致性分析

4 實驗與結果

第四章將會深入探討以生成式 AI 技術為基礎的醫療語音辨識系統構建過程及其效能。我們將先從真實病歷與生成病歷的比較著手，探索 ChatGPT 與 DaVinci 所生成的 ChaVinci 語

指標	Mes2023	DaVinci	ChatGPT
是否需要實病例的範例	否	是	是
是否需要專業性的錄音	是	否	否
製作成本	高 (需專業人員)	低 (僅模型費用)	低 (僅模型費用)
標注的成本	高 (需人工標注)	低 (無需標注)	低 (無需標注)
靈活性	低 (受限於背景知識)	中 (調整 prompt)	高 (prompt 與網路瀏覽)
多樣性	低 (表達風格相對固定)	中 (調整 prompt)	高 (prompt 與網路瀏覽)
便捷性	低 (耗時費力)	高 (可快速生成)	高 (可快速生成)
專業性	高 (具有豐富醫學知識)	中 (取決於模型)	中 (取決於模型)
資料隱私風險	中風險 (護士撰寫擬真案例)	低風險 (生成數據)	低風險 (生成數據)

表 8：醫療病例生成比較表

料庫的優勢與應用，特別是在大規模生成文本上的效率與多樣性。接著，將說明如何利用這些生成的病歷文本作為腳本，建立醫療語音資料庫，並對應到不同文字轉語音(TTS)技術的辨識性能差異。最後，我們將進行不同語料庫的 ASR 效能比較，探討 GPT 所生成的語料在靈活性、多樣性和專業性上的表現，並分析其與真實病歷和 NER 技術之間的優劣，進一步展示生成式 AI 在醫療應用中的潛力。

4.1 真實病歷與生成病歷

利用 ChatGPT-3.5 以及 DaVinci 所生成的 ChaVinci 語料庫，共含 ChatGPT 生成的 450 份病歷，以及由 DaVinci 生成的 650 份病歷組成，總計 1100 份病歷。表 8 比較了由成大醫院護理科所準備六百份病例的 Mes2023，以及 DaVinci 和 ChatGPT 所生成擬真病歷文本的各項指標。結果顯示，Mes2023 具備高度的專業性和細緻的語境捕捉能力。然而針對真實病歷採集方法的缺點是成本高昂，且撰寫過程耗時，文本的多樣性和靈活性也受到撰寫者個人表達風格和知識背景的限制。

相較而言，GPT 在靈活性、多樣性、便捷性等方面顯著優於 Mes2023。通過大量醫療數據的訓練，GPT 能夠快速生成內容豐富且語境連貫的文本，並且可以靈活調整以適應不同的情境需求。雖然在專業性上可能略遜於真人撰寫，但其生成的文本仍能達到臨床應用的要求，並且在便捷性和生成效率上具有無可比擬的優勢。這使得 GPT 成為一種極具潛力的工具，特別是在需要大規模、快速生成醫療文本的場景中。

4.2 生成 ChaVinci 為腳本的醫療語音庫

當作生成訓練 ASR 的語音庫，我們利用 ChaVinci 語料庫中所生成擬真病歷當作錄製醫療語音的腳本，然後分析相應 ASR 的辨識績效。為了快速檢驗此概念我們採用兩種不同的文字轉語音 (TTS) 技術：Google TTS 以及 Microsoft TTS，

合成相應的語音庫。總計包括 1100 份病歷的 ChaVinci 共生成，總時長約 17 個小時半的生成語音庫，如表 9。而在音訊生成部分，根據兩種不同的文字轉語音 (TTS) 服務平台音檔分割方式也不同。

語料庫	句數 (訓練 /測試)	平均 時長	平均 字數	總時長
Chavinci	6802 (5208 /1597)	9.31 s	56.16	1055.67 min (17hr 35.67 min)

表 9：Chavinci 語料庫資料分布

第一種是使用 Google TTS 服務的 ChaVinci-g，在 ChaVinci-g 中只有使用單一聲音生成音檔，從生成的語音可以明確地辨識出它是由機器生成。其語音雖然清晰，但在語調和發音上仍顯得稍微機械化，不夠自然。

第二種是使用 Microsoft TTS 服務的 ChaVinci-m，在 ChaVinci-m 中使用到了 8 個聲

	Training				Testing			
語料庫	npsChimes14	Mes2023	ChaVinci-g	ChaVinci-m	npsChimes14	Mes2023	ChaVinci-g	ChaVinci-m
病歷總數	394	499	800		101	122	300	
句子總數	5682	1573	5208		395	1573	1594	
Characters	2234	3363	2980		1940	1633	2108	
registered C /總出現數	2234/ 189181 (1.18%)	3363/ 422069 (0.8%)	2980/ 198118 (1.5%)		1807/ 39004 (4.63%)	1511/ 54305 (2.78%)	1859/60946 (3.05%)	
Keyword	973	1946	**2070**		758	699	1345	
registered K /總出現數	973/15212 (6.4%)	1946/34296 (5.67%)	2070/22049 (9.39%)		629/3179 (19.79%)	586/4675 (12.53%)	903/6786 (13.31%)	
OOV /OOV_frequency	X	X	X	X	122/183 **(66.67%)**	133/186 (71.51%)	**249**/298 (83.56%)	
OOK	X	X	X	X	**113**	129	**442**	
CER	X	X	X	X	14.3%	13.16%	**4.23%**	6.15%
KER	X	X	X	X	18.72%	18.11%	**9.61%**	13.58%
OOK_URR	X	X	X	X	40/113 (35.4%)	42/129 (32.56%)	240 /442 **(54.3%)**	210/442 (47.51%)
OOK_ORF	X	X	X	X	55/155 (35.48%)	51/160 (31.88%)	300/531 **(56.5%)**	255/527 (42.69%)

表 10：npsChimes14、Mes2023、以及 ChaVinci-g 和 ChaVinci-m 指標對比

音生成音檔，在訓練集中聲音種類有 4 女 1 男(聲音 id 為#35) ，而在測試集中有 2 女 1 男(聲音 id 為#38)，聲音種類在資料集中沒有重複，相較 Google 的 TTS，Microsoft 所生成的音檔更靠近人聲。

4.3 不同語料庫訓練的 ASR 效能比較

我們針對四個不同的語料庫進行比較，分別是 Mes2023、npsChimes14、ChaVinci-g 和 ChaVinci-m。使用 Conformer ASR 架構，在每個語料庫上進行了訓練和測試，表 10 為實驗結果。

在表 10 中，我們比較了 npsChimes14、Mes2023、與 GPT 生成的 ChaVinci 病歷文本的多樣本。其中涵蓋多項指標，**病歷總數**表示語料中包含的病歷文本的總數，而**句子總數**則是語料中所有句子的總數，這些句子以句號分割計算。**Characters** 代表語料中的唯一 token (token_unique)，並不計算出現的次數。這裡的 token 是一顆顆中文字和英文單音節。**registered C / 總出現數**表示在訓練集中已經出現的 token_unique (即 registered C)占測試集中所有 token 出現的總次數的比例。**Keyword** 是指醫療文本中的專業醫療詞彙，如藥品名稱或診斷名稱。**registered K / 總出現數** 表示在訓練集中已出現的 keyword_unique 占測試集中所有 keyword 出現總次數的比例。

OOV (Out of Vocabulary) 指訓練集中未出現、只在測試集中出現的 token_unique，而 **OOV_frequency** 指測試集中 OOV 出現的總次數。**OOK** (Out of Keyword) 則指訓練集中未出現、只在測試集中出現的關鍵詞。**OOK_URR** (OOK Unique Recognition Rate) 為在測試及中有被正確辨識出 OOK_unique/ OOK_unique 數量。**OOK_ORF** (OOK Overall Recognition Frequency) 為在測試集中有被正確辨識出 OOK 總次數/ OOK 出現的總次數。

在語音識別準確性方面，**CER** (Character Error Rate)是用來衡量字元錯誤率的指標，而 **KER** (Keyword Error Rate)則用來衡量系統對關鍵詞的辨識準確度。

我們以語料庫中涉及的單字與關鍵詞，當作衡量多元性的指標，與測試集文本中未曾出現在訓練集文本中的單字以及關鍵詞，即 OOV 與 OOK ，當作生成式工具的原創性指標。在本表中，npsChimese14 是耗時三年所搜集的病歷，其訓練集中不重複的單字為 2234，所涉及的關鍵詞有 973 個。為了擴增多元性，Mes2023 是動用 28 位成大護理師，特別安排涵蓋 16 個醫學科別的病歷，而得到不重複的單字 3363 個與關鍵詞 1946。另外 npsChimes14 與 Mes2023 測試集中，所有出現不重複的單字分別為 1940 與 1633 個，而未曾出現在訓練集中的單字 OOV 則各有 122 與 133；而測試集

的關鍵詞部份共出現 758 與 699 個，而未曾出現在訓練集中的 OOK 各別有 113 與 129。

相較而言，以 npsChimes14 為文本範例所生成的 ChaVinci 病歷，在不動用護理師的情況下，所生成病歷數與前兩個語料庫相當的情況下，訓練文本包括 2980 個單字、2070 個關鍵詞；而測試文本包括 2108 單字與 1345 個關鍵詞，而最重要的是 OOV 含 249 個，OOK 有 442。優於 Mes 所對應的 133 與 129。ChaVinci 系列的關鍵詞數量在訓練和測試集中均超過其他語料庫，顯示 GPT 能生成豐富且連貫的醫療專業詞彙。此外，ChaVinci 在 OOV 數量上也最多，表明 GPT 生成的文本具備更多樣性和新穎性，能靈活適應不同醫療情境和詞彙需求。

在醫療文本生成中，相較於真人撰寫和 NER 技術，我們可以看到 GPT 在靈活性、多樣性和專業性方面展現出了無可比擬的優勢，使其在醫療仿病歷的生成上更具競爭力。首先，從靈活與多樣性上看，GPT 生成的文本顯著優於由真人撰寫的文本與 NER 技術。真人撰寫的病歷雖然能夠準確反映專業醫療知識，但受限於個人的語言表達風格和專業背景，其生成的文本往往有一定的局限性；而 NER 技術主要依賴於預先設計好的標籤和規則，這使得它在生成新穎或未曾見過的語料時，容易受到原始數據的限制，缺乏足夠的多樣性；相比之下，GPT 能夠通過大量醫療數據的訓練，學習並模擬多種語言風格和表達方式，透過下不同的 prompt 的調整使得生成的文本能夠適應不同的語境需求，並具備更高的靈活性。此外，GPT 還能透過「網路瀏覽功能」從網路上獲取最新的醫療研究和臨床實踐資料，不斷更新和擴展其生成的內容，這是 NER 技術難以達到的。

從成本考量，真人撰寫病歷耗時且昂貴，NER 技術需標注與結構分析，過程繁瑣。而 GPT 能在短時間內生成大量不同風格的仿病歷，無需標注或微調模型。在專業性與語境捕捉上，GPT 表現更佳。真人撰寫受限於個人風格和負荷，品質不一致；NER 依賴標籤和規則，難以捕捉細微語境差異，生成文本易失去連貫性。GPT 經過大規模語料訓練，能生成更連貫、符合臨床實踐且專業用語準確的文本。

總結來說，GPT 在醫療文本生成中的靈活性、多樣性、便捷性和專業性，使其在仿病歷生成上遠優於真人撰寫和 NER 技術。GPT 不僅能夠快速生成高質量且多樣化的醫療文本，還能適應不斷變化的醫療情境，提供更準確的語境理解和專業術語應用。這些優勢使得 GPT 在醫療應用中的潛力無限，特別是在需要高效生成專業文本的場景中，GPT 的應用前景無疑是最為廣闊的。

5 結論

前述採用大語言模型的病歷生成技術可應用於更多的情境中，如以下表 11。

這些利用 GPT 用來生成病歷範例文本的來源，一共有：(1) psChiMeS14：來自衛福部台北醫院護理站合作的 516 份病歷；(2) Mes2023：來自成大醫院護理科合作的六百份病例；還有 (3) 網路上醫療網的爬文。

首先，psChiMeS14 被用來當作 DaVinci 以及 chatGPT-3.5 生成仿真病歷的範本：ChaVinci2023 (ChaVinci)。當作成功的概念驗證 (Proof of Concept) 我們的觀察是 DaVinci 生成的效果相對比較好，這可能是同屬於 chatGPT-3.0 訓練的模型，DaVinci 本就是文本產生功能中最優的，而 chatGPT-3.5 是特別針對問答應用微調。當有效果較優的文字轉語音 (Text2Speech) 模組搭配，這些生成的病歷有望用來當作錄製訓練 ASR 的腳本，以提升 ASR 的辨識效果。

其次，Mes2023 被用來當作 chatGPT-3.5 與 chatGPT-4o 生成仿真病歷的範本。主要是要生成特定科別的擬真病歷資料補足之前收集病歷僅集中在部份科別的限制，這些生成的病歷並由醫護人員進行其中 200 份生成病歷合理性的修改：Vet-Mes2023。另外，我們也在這 600 份的實際病例範例上，產生更多元性，即包括更多醫科別專業術語的病例，以作為訓練特別針對醫療情境之自然語言處理用的 Medical Bert 的訓練文本：GPT-Mes2023。此外，我們生成部份擬真病歷，實槌：「生成雙層式 NER 標註，並進一步協助雙層式模板置換」概念的不太可行性。最後，不限於我們自行收集護理交班情境的文本，我們也從網路上面所取得其他醫療網的文本，並以 chatGPT-4o 生成用來測試 ASR 的測試語音的

產出項目	生成擬真病歷目的	模型	輸入文本範例	輸出份數
Chavinci	驗證 chatGPT 生成擬真病歷文本之可行性	• DaVinci (Model: text-davinci-003) • GPT (Model: gpt-3.5-turbo-instruct)	psChiMeS14	• DaVinci:650 份 • GPT:450 份
Vet-Mes2023	生成特定醫科別的擬真病歷，並提供成大醫院護理師審核與修改	GPT (Model: gpt-3.5-turbo)	Mes2023	200 份病歷
GPT-Mes2023	生成擬真交班病歷當作訓練文本，微調醫療情境 MedCKIPBERT	GPT (Model: gpt-4o)	Mes2023	7460 句
SNOMED-CT-Mes2023	概念驗證「結合 SNOMED CT 之雙層 NER 技術嘗試」之可行性	GPT (Model: gpt-4 網頁板)	Mes2023	各 5 份
AI-Gen-Mes	生成病房交班情境外之一般醫療語音之錄音腳本	GPT (Model: gpt-4o)	一般醫療文本(網上蒐集)	354 份病歷

表 11：生成技術更多的應用情境

錄音腳本：AI-Gen-Mes。這被用來作為微調醫療情境自然語言處理用 Bert 的訓練樣本。

References

Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

LeClair, A., Jiang, S., & McMillan, C. (2019). *A neural model for generating natural language summaries of program subroutines.* Paper presented at the 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE).

Radford, A. (2018). Improving language understanding by generative pre-training.

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research, 21*(140), 1-67.

Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). *Best practices for the human evaluation of automatically generated text.* Paper presented at the Proceedings of the 12th International Conference on Natural Language Generation.

Zhang, H., Xu, J., & Wang, J. (2019). Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243*.

Selecting Better ChatGPT prompts for NLP Tasks

Aleksandra Smolka
SNHCC TIGP
Institute of Information Science
Academia Sinica
alsm@iis.sinica.edu.tw

Hsin-Min Wang
Institute of Information Science
Academia Sinica
whm@iis.sinica.edu.tw

Jason S. Chang
Department of Computer Science
National Tsing Hua University
jason@nlplab.cc

Keh-Yih Su
Institute of Information Science
Academia Sinica
kysu@iis.sinica.edu.tw

Abstract

One of the crucial parts of using ChatGPT is adopting a proper prompt to obtain the desired answer from it. However, how different prompt designs affect ChatGPT performance is still not well studied. In this work, we concentrate on three selected natural language processing (NLP) tasks (i.e., paraphrase judgment, natural language inference, and question answering), as we have found that previous works in this area lack systematic analysis of how prompts should be set. We thus propose the *prompt formula*, which is a set of binary prompt features characterizing the prompt, for systematically testing various characteristics of prompts (such as the politeness of the language, answer type, and label specification). We then identify the prompt feature set that performs best in the zero- and few-shot scenarios. The experiments show that the appropriate prompt formula could improve the ChatGPT performance by up to 15%, in comparison with those existing prompting formats reported in the NLP literature. We also show that adding training samples (i.e., the few-shot case) sometimes even deteriorates the performance.

Keywords: ChatGPT, prompt engineering, paraphrase judgment, question answering, natural language inference, LLMs

1 Introduction

ChatGPT[1] released as an API model by OpenAI is a generative foundation model specializing in language processing (Wang et al., 2023). Among those recently released Generative AI models (Gozalo-Brizuela and Garrido-Merchan, 2023), ChatGPT has gained significant popularity not only because of the commercial success of its user-end API but also due to the ease of use for a large variety of tasks (Wang et al., 2023).

Although numerous publications on ChatGPT discuss its limitations and its evaluation, we found that the impact of varying the prompt formats on performance is still under-represented in the natural language processing (NLP) literature. The lack of sufficient attention to the issue of selecting a proper *prompt pattern* in the experiment design is especially noticeable during evaluating ChatGPT on specific NLP tasks (such as paraphrase judgment (PJ), natural language inference (NLI), and question answering (QA)). In most related works, the prompt formats are either just arbitrarily set without sufficient justification in advance (Wang et al., 2023), or directly borrowed from previous works without explanation (Shen et al., 2023; Zhong et al., 2023).

As a result, the impact of using various prompt formulas is either not systematically compared or lacks in-depth analysis (Basmov et al., 2023; Kocon et al., 2023). Unfortunately, using ChatGPT without understanding the effect of prompt pattern on the model accuracy could lead to unfair performance comparison (Qin et al., 2023). Since

[1] https://openai.com/chatgpt/

the training data/procedure of ChatGPT is not transparent to the end-user, and it has been shown that *prompt engineering* considerably affects the performance (Qin et al., 2023), it is thus important to test and get an appropriate prompt pattern before ChatGPT is used for the given task.

To achieve the above goal, we first select features re-occurring in the prompts used by previous works (Basmov et al., 2023; Jang and Lukasiewicz, 2023; Lai et al., 2023, Shen et al., 2023; Wang et al., 2023), and then propose the *prompt formula*, which is a set of binary prompt features (i.e., "+" or "-", which denote *on* and *off* states, respectively) that characterize the prompt. For example, the usage of polite words could be adopted to categorize various prompts, as shown in Figure 1 and Table 1. Afterward, we use this prompt formula to systematically select the appropriate prompt for the given task, which is done via the Sequential-Forward-Selection (SFS; Ferri et al., 1994) feature-set selection procedure. An appropriate prompt formula would be found for each specific NLP task.

To show the superiority of the proposed approach, we design a series of experiments to check how those features impact the model results. For each task, we manually create its associated *base prompt*, which is a common prompt additionally interleaved with a few pre-specified empty placeholders. Each placeholder is used to insert additional information associated with a specific feature. Figure 1 shows an example of the feature "*List allowable answers*" (under the "*feat.*" column in the right table), which has been filled with the corresponding purple string *<reply "paraphrased" or "not-paraphrased">* (left).

Each type of placeholder corresponds to a specific prompt feature. Furthermore, we propose a *data augmentation method*, which instantiates the above base prompt by inserting the corresponding content into those placeholders. In this way, we can generate the corresponding prompt, which will be applied to a specific NLP task later, for each specified feature configuration. We then adopt the SFS approach to select the best feature configuration (among 6 different features) for each adopted NLP task in both zero- and few-shot scenarios.

Our experiments adopt several types of datasets: (1) two PJ datasets including MRPC (Dolan and Brockett, 2005) and QQP (Aghaebrahimian, 2017), (2) an NLI dataset (SNLI; Bowman et al., 2015), and (3) a multiple-choice QA dataset (CommonsenseQA; Talmor et al., 2019). We select these datasets as they are frequently adopted to test common NLP tasks with varying levels of difficulty.

Based on our experiments, we show that carefully selecting an appropriate prompt formula could improve the results by up to 15% in comparison with the best prompt formulas adopted in previous representative works. We further show that the current ChatGPT model with GPT-3.5 backbone sometimes lacks the ability to generalize

Please answer whether S1 and S2 are paraphrased or not. You should consider the syntax and semantics of the sentences to compare their meaning. Please reply "paraphrased" or "not-paraphrased". Your answer should be only one word, in lowercase letters.
S1: [...]
S2: [...]
Answer:

feat. \ con.	Content filled in the placeholder (+)	Content filled in the placeholder (-)	ex.
polite words	"please"	∅	+
state desired model action	"answer"	∅	+
way to integrate test sample	integrated within the prompt text	cited below using variables	-
list key competencies	"...consider the syntax..."	∅	+
list allowable answers	"reply "paraphrased" or ..."	∅	+
specify answer format	"your answer should..."	∅	+

Figure 1: An example for instantiated prompt (left block) used for the paraphrase judgement task (its associated *prompt formula* and *base prompt* are shown in the right part). Each color indicates a specific kind of placeholder that has been filled with the corresponding content of the associated feature value (as listed in the table). On the right table, the "*feat.*" column indicates the corresponding feature names; the "*Con.*" column is the content for the corresponding binary feature value, and the "*ex.*" column is the corresponding binary feature value (i.e., "+" and "-") for the prompt example given on the left.

from a small number of examples (i.e., the few-shot scenario) to other similar cases.

Our contributions include:

- Analyzing all the prompts collected from those published NLP-related papers that we have read (mainly related to PJ, NLI, and QA), and then categorizing them into 6 different binary features.
- Proposing the *prompt formula* to use it for systematically selecting the appropriate prompt for the given task.
- Using SFS to look for the best corresponding prompt formula for each task, which allows up to 15% improvement on selected tasks in comparison with the best prompting method listed in the related literature.
- Showing that adding examples to the prompt sometimes even deteriorates the performance of the few-shot scenario in selected tasks, which violates the common intuition.
- Releasing the code, which can be used to select the best corresponding prompt formulas for other datasets/tasks.[2]

2 Feature selection and prompt generation

In this section, we first categorize the prompt patterns with distinctive features (Section 2.1) and then show how we use these features to generate the prompts that could be used in our experiments (Section 2.2).

2.1 Prompt pattern categorization

We begin by analyzing a range of existing prompts (concentrating on those applied to PJ, NLI, and QA tasks[3]). We then categorize them into 4 main prompt categories (as shown in Table 1) depending on: (1) adopting polite words such as "*please*" (Lai et al., 2023); (2) attaching the information about how to solve the problem (Liu et al., 2023a; Wang et al., 2023); (3) attaching the information about the desired answer (Basmov et al., 2023); (4) adopting a specific way for integrating the test sample into the prompt (i.e., whether the test sample is inserted within the prompt text or below (Shen et al., 2023)). Each category will be further elaborated as follows.

The first category of adopting polite words is rather self-explanatory and usually involves adding words such as "*please*" or using politer modal verbs (e.g., "*could*" instead of "*can*") (Lai et al., 2023). The second main category includes two

Work	task	Adopt polite words	Attach information about solving the problem		Attaching information about the desired answer		Way to integrate the test sample
			Desired action	List key competencies	Allowable answers	Answer format	
Jang and Lukasiewicz (2023)	PJ	-	-	-	-	-	-
Wang et al. (2023)	PJ	-	+	-	+	-	-
Basmov et al. (2023)	NLI	-	+	+	+	+	-
Lai et al. (2023)	NLI	+	+	-	+	+	-
	QA	-	+	+	+	+	-
Shen et al., (2023)	QA	-	+	-	+	-	+

Table 1: Prompt features in previous representative works. The "+" sign indicates that the feature is present. In contrast, the "-" sign indicates that it is not. The prompt formulas vary greatly across different works, and even among the tasks mentioned within one publication.

[2] https://github.com/alsmolka/gpt-prompt-analysis

[3] In our work, we concentrate only on single-stage prompting such as that mentioned in Lai et al. (2023), and do not consider multiple-stage prompting such as that adopted in Kojima et al., (2022) or Qin et al., (2023). However, our proposed approach can be also applied to the multiple-stage prompting with slight modification, which is beyond the scope of this work.

sub-categories describing whether the prompt provides additional information for solving the problem. The prompt might either: (a) explicitly include a verb describing the desired action to get the answer (e.g., "*select*", "*compare*", … (Wang et al., 2023)), or (b) list key competencies, which explains what type of skill or information is needed to solve the problem (e.g., "*use commonsense knowledge*" (Lai et al., 2023)).

The third category specifies whether the desired answer is explicitly specified within the prompt. It also includes two sub-categories, and can either: (a) list all allowable answers (e.g., "*yes*" or "*no*" for binary classification (Wang et al., 2023)), or (b) specify the desired answer format (e.g., limiting the desired answer to have only one word, (Basmov et al., 2023)).

Finally, the last way of grouping the prompts is based on how the test sample is integrated into the prompt: (1) directly cited within the prompt text (e.g., *"Are following sentences paraphrases: [sentence1], [sentence2]"* (Shen et al., 2023); corresponding to the "+" sign of the feature "*way to integrate the test sample*" in both Figure 1 and Table1), or (2) provided below the main prompt in which it is mentioned as a variable (e.g., "*Are S1 and S2 paraphrases*", with *S1* and *S2* specified below the whole prompt (Lai et al., 2023); corresponding to the "-" sign and adopted in Figure 1).

task	work	Test dataset	# SPL	Acc.
PJ	Jang and Lukasiewicz, (2023)	MRPC	1000	0.54
	Wang et al., (2023)	**MRPC**	**1000**	**0.60**
	Jang and Lukasiewicz, (2023)	**QQP**	**1000**	**0.72**
	Wang et al., (2023)	QQP	1000	0.50
NLI	**Basmov et al. (2023)**	**SNLI**	**1000**	**0.47**
	Lai et al. (2023)	SNLI	1000	0.45
QA	Lai et al., (2023)	CQA	1000	0.57
	Shen et al., (2023)	**CQA**	**1000**	**0.61**

Table 2: Performance of test prompt formulas from two representative previous works ("work") for each task ("task") and dataset ("test dataset"). Each row corresponds to a single formula. #SPL indicates the number of prompted samples, and "Acc" indicates the associated accuracy using the given prompts. Selected baselines are in bold.

Table 1 shows these adopted binary features together with the corresponding contents adopted in two previous representative works (for each task). Each of the features can be specified with a binary value (i.e., +/-).

2.2 Prompt generation for specific tasks

To better specify and systematically evaluate different prompts, we propose using a pre-specified *prompt formula*, which is a configuration consisting of all 6[4] binary feature values mentioned above, to generate various prompts with a specific set of feature values (Table 1). We also propose a *base prompt*, which contains various placeholders of several types as shown in Figure 1 (colored boxes), to use for generating the prompts used in ChatGPT. We fill these placeholders according to the assigned feature values specified in the corresponding prompt formula. For example, for the feature "*polite words*", the two pink boxes in Figure 1 will be filled with the word "*please*" if the associated feature value is positive.

We manually specify a specific base prompt for each task, in which additional text would be inserted into the corresponding placeholders (associated with various features) to generate the desired prompt. In the few-shot scenario, the additional benchmark examples are simply inserted below the zero-shot base prompt.

3 Experiments

3.1 Datasets

We select our benchmark datasets across a range of NLP tasks, including PJ (i.e., MRPC (Dolan and Brockett, 2005) and QQP (Aghaebrahimian, 2017)), NLI (i.e., SNLI (Bowman et al., 2015)), and multiple-answer QA (i.e., CommonsenseQA (Talmor et al., 2019)).

To prepare the test set for each task listed above, we randomly take 1,000 samples for the

[4] Each sub-category listed in Section 2.1 is counted as a distinct feature.

development and 1,000 samples for the test set, keeping the class balance.[5] We repeat sampling for each task and dataset. We then generate the prompts for the ChatGPT model as described in Section 2.2. The development set in this scenario is used to select the best feature configuration (i.e., the best prompt formula) and the test set is used to report the final results.

For the few-shot scenario experiments (i.e., Experiment 2), we take the best prompt formula found in Experiment 1 for each task and follow Liu et al. (2023a) to additionally augment it with some randomly sampled training data. We test 4 variations for adding the training data into the prompts, including (a) 1 random example, (b) 2 examples randomly sampled regardless of their classes (for all benchmarks), (c) 2 random examples each from a different class, and (d) 4 random examples, two from each class. The last two variants are only conducted for the tasks for which benchmark datasets are associated with 2 classes (i.e., PJ-MRPC and PJ-QQP).

Furthermore, to measure the susceptibility of the model to the change in the training sample provided in the prompt, we calculate the standard deviation (SD) of its performance (Sekander Hayat Khan, 2011) on 5 sets of 1,000 generated answers. Each set uses different training samples. It thus gives us a total of 60K prompted samples to be used in Experiment 2.

3.2 Experimental settings

In our experiments, we adopt the ChatGPT API with the GPT-3.5-turbo model. Following the approach adopted in the previous publications (e.g., Wang et al., 2023), the model is not fine-tuned. Hence, the same model is always used in either the zero-shot (i.e., baseline (Section 3.3), and Experiment 1 (Section 3.4.1)) or the few-shot scenario (i.e., Experiment 2 (Section 3.4.2)).

To simplify the performance evaluation step, the answers obtained from ChatGPT are first automatically normalized before evaluation, which is similar to what has been done in the previous works (Basmov et al., 2023), as we have found that the format inconsistency between the benchmark and the obtained answer causes a huge drop in its performance. The normalization procedure is thus introduced to reduce the output format variation and includes the following three procedures: (1) removing extra whitespaces, (2) correcting punctuation, which involves removing unnecessary punctuation marks, and (3) casting all words to lowercase and removing surplus strings (e.g., the output *"S1 and S2 are paraphrases"* would be automatically converted into the benchmark label *"paraphrase"*).

Feat. name		Development set							Test set	
		none	plt.	act.	Lbl.	fmt.	cmpt.	int.	baseline	BEST
Task-dataset	PJ-MRPC	0.54	0.51	0.63	0.56	0.52	0.60	0.62	0.60	**0.72***
	PJ-QQP	0.72	0.71	0.60	0.37	0.64	0.49	0.69	0.69	0.69
	NLI-SNLI	0.39	0.19	0.44	0.43	0.49	0.11	0.38	0.46	**0.61***
	QA-CQA	0.45	0.43	0.61	0.49	0.60	0.38	0.46	0.59	0.59

Table 3: Selection of the best prompt formula via the sequential-forward-selection (SFS; Ferri et al., 1994) procedure (the zero-shot scenario). We report the accuracy measured from the following cases: (1) none of the feature-switch is turned on (the first *none* column, treated as an additional pseudo-feature in the table), (2) only one feature-switch is activated (the columns 2-7), (3) the baseline prompt from Table 2 (the *baseline* column), and (4) the best prompt formula found via the SFS feature-set selection procedure (the "*BEST*" column). The features used in the experiment: "*plt.*" – using the polite words; "*act*" – specifying the required action; "*Lbl.*" – specifying the allowable answer labels; "*fmt*" – specifying the answer format; "*cmpt.*" – outlining the competencies needed; "*int.*" – integrating the test sample into the prompt text. Results showing statistically significant improvement over the corresponding baseline ($p<0.05$) are marked with a "*" symbol.

[5] This method of preparing data for ChatGPT evaluation is similar to those previous approaches such as in Shen et al. (2023).

Task-dataset	Prompt formula	Adopt polite language	Attach information about solving the problem		Attaching information about the desired answer		Way to integrate the test sample
			Desired action	List key competencies	Allowable answers	Answer format	
PJ-MRPC	baseline	-	+	-	+	-	-
	ours	+	+	-	-	-	+
PJ-QQP	baseline	-	-	-	-	-	-
	ours	-	-	-	-	-	-
NLI-SNLI	baseline	-	+	+	+	+	-
	ours	-	+	-	-	+	+
QA-CQA	baseline	-	+	-	+	-	+
	ours	-	+	+	-	+	-

Table 4. Comparison of features between our best prompt and the baseline for each task-dataset combination in the zero-shot scenario. The feature names (columns) are the same as those in Table 1.

3.3 Baseline selection

We test each task-dataset pair using two different formulas taken from earlier representative works as listed in Table 1. Since it is unclear which previous approach yields the best results based only on the literature review, we choose the prompt formula with the best performance on our test set as its baseline. As a result, we end up with a total of four baselines (i.e., one for each task-dataset combination).

Table 2 shows the performance of the prompt formulas tested across the task-dataset pairs. Prompts selected as baselines are marked in bold. The table additionally shows that the performance of adopting various prompt formulas can vary greatly even on the same task-dataset combination (e.g., up to 22%, from 0.50 to 0.72, on the PJ-QQP dataset), highlighting the importance of selecting an appropriate prompt formula while using ChatGPT.

3.4 Experimental results

After establishing the baselines for performance comparison, we conduct two key experiments to test the impact of each prompt feature described in previous sections. In Experiment 1 (Section 3.4.1), we identify the best prompt formula based on the prompt features outlined in Section 2, focusing on the zero-shot scenario. Next, in Experiment 2, we check whether adding a few training examples to the prompt (i.e., a few-shot scenario) can enhance the model performance (Section 3.4.2).

3.4.1 Experiment 1 (Zero-Shot Scenario)

In the first experiment, we adopt the zero-shot scenario. We evaluate the effect of each feature we have identified during prompt analysis (Section 2.2, listed in Table 1). Afterward, we apply the SFS (Sequential Forward Selection) procedure (Ferri et al., 1994) to select the best prompt as represented by a feature set. The experiment is performed individually for each task-dataset combination and evaluated on the development set. The SPF approach allows us to identify the best feature combination by progressively adding individual features to the feature set greedily. At each selection step, we check which feature addition results in the best accuracy improvement. The process continues until the potential feature set is exhausted or if adding more features does not yield any further improvements. The best feature set at this stage becomes our final set of features.

Table 3 reports the accuracy for various prompt formulas: (1) with all features turned off (on the development set), (2) individually activating each feature (on the development set), (3) the baseline prompts selected from Table 2 (on the test set), and (4) the best prompt formulas found via the SFS procedure (on the test set).

To compare the established baselines (from Section 3.3) with the best prompt formulas obtained above, we apply the Student's t-test (Student, 1992). We uniformly split the whole test set into *n* subsets (*n=10*) and calculate the *mean* and *variance* of the model's accuracy on these subsets. Results showing statistically significant improvement over the corresponding baseline ($p<0.05$) are marked with an asterisk ("*").

Table 3 shows that the best prompt formula (the associated feature setting is given in Table 4) can outperform its corresponding baseline by up to 15% (NLI-SNLI, 46% vs. 61%). For the PJ-MRPC task, the improvement is 12% (60% vs. 72%). In the remaining two datasets (i.e., PJ-QQP and QA-CQA), our best prompts also match the performance of the best prompt used in two previous representative works (i.e., the baseline).

Table 4 compares the feature-setting of each best prompt formula selected with that of its baseline. It seems that each feature could be beneficial for some task-dataset combinations, as all of them appear somewhere in the best feature configurations. Interestingly, for certain task-dataset combinations, it is most beneficial to have all the features turned off. This is the case for the PJ-QQP task, where the prompt formula without activating any features proved to be the best (accuracy of 69% for both the baseline and our approach; see the PJ-QQP row in Table 3). It might be because the QQP dataset is noisy.[6] Providing additional information about the answer (e.g., the answer format) to ChatGPT might confuse, as the benchmark samples do not always follow the desired form specified in the prompt.

While it is difficult to draw any general conclusion from Table 4, it seems that explicitly stating the expected action for getting the answer (by using verbs such as "*select the answer*" or "*compare*") is usually beneficial to the model performance. This is supported by the observation

```
Please answer whether S1 and S2 are paraphrased or
not.
Example:
S1: "They are trying to turn him into a martyr, said
Vicki Saporta, president of the National Abortion
Federation, which tracks abortion-related violence."
S2: "We need to take these threats seriously, said
Vicki Saporta, president of the National Abortion
Federation."
Answer: negative
S1: "Mahmud controlled access to Saddam for everyone
but immediate family members, Pentagon officials
said."
S2: "Mahmud controlled access to Saddam and was
frequently at his side."
Answer:
```

Figure 2. A few-shot prompt example for the PJ-MRPC task with a single negative training example added.

that the "*Desired action*" feature is activated in most of the best prompt formulas.

On the other hand, not adopting polite wording gives better results in all but one case (i.e., our best PJ-MRPC prompt), and is also not adopted in most related works (as shown by Table 1). We guess that the majority of training data might not adopt the

Task-dataset	Few-shot (SD)				Zero-shot
	1	1P+1N	2	2P+2N	
PJ-MRPC	0.64 (0.02)	0.69 (0.01)	0.67 (0.02)	0.66 (0.03)	**0.72**
PJ-QQP	0.39 (0.13)	0.71 (0.04)	0.62 (0.09)	**0.76** (0.03)	0.69
NLI-SNLI	0.57 (0.03)	n/a	**0.64** (0.01)	n/a	0.61
QA-CQA	0.45 (0.03)	n/a	**0.69** (0.01)	n/a	0.59

Table 5. Comparison of zero-shot and few-shot results using the best prompt selected in Experiment 1. The columns "1", "1P+1N", "2" and "2P+2N" indicate various numbers of added training examples (P-positive, N-negative). "n/a" cells denote the datasets with more than 2 classes; as a result, indicated scenarios cannot be applied.

polite wording; however, it cannot be confirmed without inspecting the ChatGPT training samples.

Similarly, providing the competencies needed to solve the problem often appears unnecessary or can even negatively impact the performance (75% of the cells under the "*List key competencies*" column have a "-" sign). It is conjectured that providing such hints might work better in the multiple-stage prompting case, in which it is usually adopted (e.g., Shi et al., 2022), rather than the single-stage prompting case like in our work.

For other features, it seems whether they should be turned on or off depends on each specific task and dataset. Overall, the experiment results (Table 3) show that our proposed prompt formula selection approach can obtain the prompts that outperform (or at least are on a par with) those adopted in previous related works.

3.4.2 Experiment 2 (Few-Shot Scenario)

In the second experiment, we check the effect of adding training examples to the prompt (i.e., the few-shot scenario). We follow Liu et al. (2023a) to add a few examples and use the best prompt formula selected in Experiment 1 to generate the new prompts by adding a few examples from the respective training set to the original prompt (*cf.* Section 3.1 for detailed explanation of how the training samples are added). Figure 2 shows a

[6] It is noisy because a large subset of it has not been manually checked after it was automatically collected from an online forum (https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs).

prompt example for the PJ-MRPC task when adding one training sample.

Table 5 shows the few-shot performance under the scenarios mentioned above. We observe that adding training samples improves the performance in three out of four tasks (i.e., PJ-QQP, NLI-SNLI, and QA-CQA). The highest improvement is 10% (QA-CQA; from 0.59 in zero-shot to 0.69 in few-shot, SD=0.01).

Interestingly, regardless of the number of added examples, the performance on the PJ-MRPC task decreases, which contradicts the common intuition. Moreover, contrary to the other cases, the performance when four examples are added is even lower than the case when only two training examples are added (i.e., 0.66 vs. 0.69; for the 2P+2N and 1P+1N cases). It is conjectured that this performance drop might be due to the high variability of MRPC samples; as a result, the added training examples could be quite different from the test sample in a given prompt, which might have a negative impact on the inference procedure.

To sum up, adding training examples to the prompt could be beneficial for getting better performance in most cases. But just like other prompt features, the result depends heavily on the benchmark dataset and the task. Hence, before adopting the few-shot scenario, it is necessary to first test if it is indeed beneficial.

3.5 Ablation test and error analysis

To show that the post-processing steps taken when obtaining answers from ChatGPT are essential, we examine the performance in Experiment 1 without conducting the automatic answer normalization. The results show that the performance of each case drops considerably, and is even close to 0 for the PJ task (on both the MRPC and QQP datasets) because returning human-acceptable un-normalized answers will not be counted as correct during automatic evaluation (e.g., generating "*Paraphrase*" or "*yes, paraphrase*" instead of the expected answer "*paraphrase*"). It thus shows that using the simple normalization procedure mentioned in Section 3.2 could greatly alleviate the format inconsistency problem.

We then check the errors made in the zero-shot scenario (i.e., Experiment 1, Section 3.4.1). We manually analyze one hundred errors randomly sampled from each benchmark dataset. Based on our analysis, the errors could be grouped into two main categories: (1) incorrect answers (overall 91%), in which ChatGPT returns wrong, but legal outputs (e.g., "*positive*" instead of "*negative*" for PJ; other tasks also behave similarly); (2) illegal answers (overall 9%), in which ChatGPT generates illegal outputs that do not match any of the allowable benchmark labels in the given dataset. This shows that although it is difficult to control the model output format, it is alleviated to some degree by our automatic normalization. Still, 9% of answers could not be recovered.

For the few-shot scenario (Experiment 2) we randomly sample one hundred errors from each benchmark dataset. Similar to the zero-shot case, most answers are legal but incorrect (94% of the cases). Interestingly, the overall percentage of illegal answers decreases by 3% in the few-shot scenario (from 9% to 6%, compared with the zero-shot case). It is conjectured that adding training examples might help the generative model (such as ChatGPT) better recognize what is the desired output format.

4 Related work

Some recent works aim to evaluate ChatGPT, especially in the context of its applicability to various NLP tasks as a task-independent model (Srivastava et al., 2022; Bang et al., 2023; Guo et al., 2023; Kocon et al., 2023). Works that concentrate exclusively on specific NLP tasks are also common, including QA (Lai et al., 2023), machine translation (Peng et al., 2023), or summarization (Goyal et al., 2022, Zhang et al., 2023).

Most works on prompt evaluation focus on multi-stage prompting (Shi et al., 2022, Wei et al., 2022, While et al., 2023, Zhou et al., 2022) as opposed to our single-stage approach, though some works are also applicable to single-stage prompting as well (e.g., Liu et al., 2023a; Zuccon and Koopman, 2023). In comparison to those works, we specifically categorize the characteristics of various prompts and propose to use the prompt formula to systematically and automatically select the best prompt feature combination for the given task.

5 Conclusions

Given the variation of prompts considerably affects the performance of ChatGPT, we propose a *prompt formula* for systematically selecting an appropriate prompt for specific tasks. We conduct prompt engineering on three NLP tasks: PJ, NLI, and QA. The experiments have shown that the prompt feature set selected via the SFS procedure could improve the performance by up to 15% in comparison with the best formula found from the selected previous works (under the zero-shot scenario). Furthermore, we also show that adding a few training examples to the prompt sometimes might even deteriorate the performance, which is contradictory to the common intuition. This underscores the importance of selecting a proper prompt formula (both in terms of content and structure) before applying ChatGPT to the downstream task. In the future, the work in this paper can be extended to include an automatic pipeline allowing for automatic prompt selection for any downstream NLP task.

References

Ahmad Aghaebrahimian. 2017. Quora Question Answer Dataset. *Text, Speech, and Dialogue.* Springer International Publishing, pages 66–73. https://doi.org/10.1007/978-3-319-64206-2_8

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.ijcnlp-main.45

Victoria Basmov, Yoav Goldberg, Reut Tsarfaty. 2023. ChatGPT and Simple Linguistic Inferences: Blind Spots and Blinds. *arXiv:2305.14785* https://doi.org/10.48550/arXiv.2305.14785

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1075

William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. https://aclanthology.org/I05-5002

F.J. Ferri, P. Pudil, M. Hatef, J. Kittler. 1994. Comparative Study of Techniques for Large-Scale Feature Selection. *Machine Intelligence and Pattern Recognition, North-Holland, Volume 16, 1994*, pages 403-413, ISSN 0923-0459, ISBN 9780444818928. https://doi.org/10.1016/B978-0-444-81892-8.50040-7.

Ronald A. Fisher. 1921. Studies in Crop Variation. I. An Examination of the Yield of Dressed Grain from Broadbalk. *Journal of Agricultural Science. 11 (2),* pages 107–135. doi:10.1017/S0021859600003750

Tanya Goyal, Junyi Jessy Li, Greg Durrett. 2022. News Summarization and Evaluation in the Era of GPT-3. *arXiv:2209.12356* https://doi.org/10.48550/arXiv.2209.12356

Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. 2023. ChatGPT is Not All You Need. A State of the Art Review of Large Generative AI Models. *arXiv:2301.04655* https://doi.org/10.48550/arXiv.2301.04655

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding Jianwei Yue, and Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv:2301.07597* https://doi.org/10.48550/arXiv.2301.07597

Myeongjun Jang and Thomas Lukasiewicz. 2023. Consistency Analysis of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* pages 15970–15985, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.991

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, Przemysław Kazienko. 2023. ChatGPT: Jack of All Trades, Master of None. *Information Fusion, Volume 99, 2023*, 101861,ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2023.101861.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-shot Reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, Article 1613, pages 22199–22213, Curran Associates Inc., Red Hook, NY, USA. https://dl.acm.org/doi/10.5555/3600270.3601883

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.878

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, Yue Zhang. 2023a. Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4. *arXiv:2304.03439* https://doi.org/10.48550/arXiv.2304.03439

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv*. 55, 9, Article 195 (September 2023), pages 1–35. https://doi.org/10.1145/3560815

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards Making the Most of ChatGPT for Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.373

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.85

M. Sekander Hayat Khan. 2011. Standard Deviation. In: Miodrag Lovric, (eds) *International Encyclopedia of Statistical Science.* Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_535

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT. *arXiv:2304.08979* https://doi.org/10.48550/arXiv.2304.08979

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, Jason Wei. 2022. Language Models are Multilingual Chain-of-Thought Reasoners. *arXiv:2210.03057* https://doi.org/10.48550/arXiv.2210.03057

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso et al. 2022. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *arXiv:2206.04615* https://doi.org/10.48550/arXiv.2206.04615

Student. 1992. The Probable Error of a Mean. In: Kotz, S., Johnson, N.L. (eds) Breakthroughs in Statistics. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-4380-9_4

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1421

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, Binxing Jiao, Yue Zhang and Xing Xie. 2023. On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. *arXiv.2302.12095.* https://doi.org/10.48550/arXiv.2302.12095

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22),* pages 24824–24837, Curran Associates Inc., Red Hook, NY, USA, Article 1800. https://dl.acm.org/doi/10.1145/3626772.3657788

Haopeng Zhang, Xiao Liu, Jiawei Zhang. 2023. Extractive Summarization via ChatGPT for Faithful Summary Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3270–3278, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.214

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, Dacheng Tao. 2023. Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT. *arXiv:2302.10198* https://doi.org/10.48550/arXiv.2302.10198

Guido Zuccon and Bevan Koopman. 2023. Dr ChatGPT, Tell Me What I Want to Hear: How Prompt Knowledge Impacts Health Answer Correctness. *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pages 15012–15022, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.928

Decremental Learning for Domain Adaptation in Neural Machine Translation

Kai Jin
School of Foreign Languages
Qilu University of Technology
250353, Jinan, Shandong, China
jk@qlu.edu.cn

Wuying Liu(✉)
Shandong Key Laboratory of Language Resources Development and Application
Ludong University
264025, Yantai, Shandong, China
wyliu@ldu.edu.cn

Abstract

Domain adaptation has become a research hotspot in neural machine translation (NMT). Based on the quantitative analysis results of two adaptabilities, we propose a two-phase decremental learning framework for scenes involving large-scale common bilingual parallel sentences and small-scale monolingual domain texts. In the domain filtering phase, we filter common sentence pairs with domain texts and train two domain NMT models with these domain sentence pairs. In the quality filtering phase, we use the trained domain NMT models to translate the domain sentence pairs and evaluate the translation quality to delete low-quality domain sentence pairs to get high-quality ones. With these high-quality sentence pairs, we train optimized bidirectional domain NMT models adapted to the specific domain. The experimental results of English-Chinese bidirectional NMT in the legal domain show that when the number of training steps decreases from 1,714,250 to 446,100, the BLEU value of the English-Chinese NMT model increases from 45.41 to 47.88, and that of the Chinese-English NMT model increases from 32.03 to 34.12, which demonstrates that the decremental learning is effective in achieving state-of-the-art performance with greatly reduced training space-time costs.

Keywords: Domain Adaptation, Domain Filtering, Quality Filtering, Bidirectional Legal NMT

1 Introduction

Recently, resource-rich general neural machine translation (NMT) has been well studied, resulting in myriad brilliant algorithms, data resources, and practical tools (Tan et al., 2020). With the explosive growth of language data, resource-rich machine translation (MT) research is paying more attention to transfer learning methods. Domain adaptation is a method of transfer learning featuring the same tasks and different domains. In our case, both tasks are MT tasks but the source domain and the target domain are different – the training set is a wide-domain English-Chinese corpus while the test set is a legal domain English-Chinese corpus. Different domains in domain adaptation can be embodied in that there is inconsistent data distribution between the source and target domains, or that there are a large number of labeled out-domain samples while a minimal number of or no labeled in-domain samples.

There are two main approaches to the study of domain adaptation NMT, which are evolved and developed from the study of domain adaptation statistical MT (Chu and Wang, 2018). One is model centric approach, and the other is data centric approach. There is an overlap between the two approaches since the former one may also use monolingual or parallel sentence data.

The model centric approach focuses on improving algorithms of neural networks. (1) Intervening in the architecture (Domhan and Hieber, 2017) (Britz et al., 2017) (Kobus et al., 2017). (2) Intervening in the training (Chen et al., 2017) (Wang et al., 2018) (Varga, 2017) (Dakwale and Monz, 2017) (Dou et al., 2019) (Chu et al., 2017) (Barone et al., 2017). (3) Intervening in the decoding (Adams et al., 2022) (Freitag and Al-Onaizan, 2016) (Khayrallah et al., 2017).

The data centric approach is more suitable for engineering applications in a shorter time. (1) By utilizing in-domain monolingual data (Currey et al., 2017) (Zhang and Zong, 2016) (Cheng et al., 2016). (2) By utilizing out-domain high-quality parallel data (Wang et al., 2017) (Wees et al., 2017). (3) Utilizing parallel data of unknown quality (Saunders, 2022) (Hu et al., 2019).

The research of domain adaptation MT aims for better performance of in-domain MT models by using information-rich out-domain samples. When there is an inconsistent distribution between the data of the training set and the test set, the trained model resulting from machine learning often overfits the source domain, thereby reducing the generalizability in the target domain. An ideal of domain adaptation MT should produce with higher efficiency a well-performing MT model corresponding to the domain. The data in the above data centric approaches are very close to real data environments. How to use this kind of parallel data of unknown quality to achieve NMT suitable for specific domains is a more specific and practical research issue. For this problem, we made quantitative analysis on two adaptabilities and proposed a novel decremental learning idea.

2 Examination of Adaptabilities

To calculate the adaptability between source domain and target domain, we first calculate the union V={v_1, v_2, …, v_n} of the source domain data vocabulary Vs and the target domain data vocabulary Vt. Then, basing on the n-dimensional vector base V, we calculate the word frequency vector S=<s_1, s_2, ..., s_n> of the source domain data and the word frequency vector T=<t_1, t_2, ..., t_n> of the target domain data. Finally, we measure the difference between the source domain and the target domain by Kullback-Leibler Divergence (KLD) and Maximum Mean Discrepancy (MMD). The KLD is widely used in adaptation machine learning tasks as a loss function (Nguyen et al., 2022). The MMD is mainly used to measure the distance between two different but related distributions (Wang et al., 2020).

We calculate the KLD and MMD of sentence pairs between the source domain and the target domain of language A, the KLD and MMD of sentence pairs between the source domain and the target domain of language B, and the KLD and MMD of parallel sentence pairs between the source domain and the target domain. We make a statistical analysis of three sets (LAW07, LAW08 and LAW09) of parallel sentences in both English and Chinese. LAW07 contains 21,942,400 pairs of sentences in the source domain, LAW08 contains 5,899,520 pairs of sentences in the source domain, and LAW09 contains 5,710,080 pairs of sentences in the source domain. These three sets all contain the same 50,000 pairs of sentences in the target domain. The English-Chinese adaptabilities are shown in Table 1. As the absolute values of KLD and MMD are very small, they are multiplied by 10^6 to be shown in the table. The values in Table 1 show that the source domain of LAW09 is the closest to the target domain, whether from the monolingual perspective of English and Chinese or from the English-Chinese bilingual perspective.

Source (number of sentences)	Target (number of sentences)	KLD×10^6	MMD×10^6
LAW07.train.eng (21,942,400)	LAW07.test.eng (50,000)	11.028	21.935
LAW08.train.eng (5,899,520)	LAW08.test.eng (50,000)	1.144	4.292
LAW09.train.eng (5,710,080)	LAW09.test.eng (50,000)	**1.098**	**2.384**
LAW07.train.zho (21,942,400)	LAW07.test.zho (50,000)	93.858	262.260
LAW08.train.zho (5,899,520)	LAW08.test.zho (50,000)	9.254	99.182
LAW09.train.zho (5,710,080)	LAW09.test.zho (50,000)	**8.911**	**69.141**
LAW07.train.engzho (43,884,800)	LAW07.test.engzho (100,000)	41.145	30.041
LAW08.train.engzho (11,799,040)	LAW08.test.engzho (100,000)	4.695	20.027
LAW09.train.engzho (11,420,160)	LAW09.test.engzho (100,000)	**4.297**	**19.550**

Table 1: Adaptabilities of English-Chinese Corpus

For the above three sets of corpus, we also calculate the Chinese-English adaptabilities, and the results are shown in Table 2. We compare the values in Table 1 and Table 2 and find that among the three groups of values the results of Chinese monolingual data are exactly the same, and there is no difference between the relative size relationship of English monolingual data and Chinese-English bilingual data, in spite of their different absolute values. From this result we can draw the same conclusion on adaptability of Chinese-English translation as that of English-Chinese translation. The difference mentioned above in Table 1 and Table 2 is due to the fact that all the letters in the English sentences in English-Chinese translation corpus are lowercase, and those in Chinese-English translation corpus follow the rule of capitalization in English, while, on the other hand, there is no such difference of form in Chinese.

Source (number of sentences)	Target (number of sentences)	KLD×10^6	MMD×10^6
LAW07.train.zho (21,942,400)	LAW07.test.zho (50,000)	93.858	262.260
LAW08.train.zho (5,899,520)	LAW08.test.zho (50,000)	9.254	99.182
LAW09.train.zho (5,710,080)	LAW09.test.zho (50,000)	**8.911**	**69.141**
LAW07.train.eng (21,942,400)	LAW07.test.eng (50,000)	11.502	24.319
LAW08.train.eng (5,899,520)	LAW08.test.eng (50,000)	1.510	4.768
LAW09.train.eng (5,710,080)	LAW09.test.eng (50,000)	**1.418**	**2.452**
LAW07.train.zhoeng (43,884,800)	LAW07.test.zhoeng (100,000)	49.440	25.272
LAW08.train.zhoeng (11,799,040)	LAW08.test.zhoeng (100,000)	7.295	15.736
LAW09.train.zhoeng (11,420,160)	LAW09.test.zhoeng (100,000)	**6.455**	**14.440**

Table 2: Adaptabilities of Chinese-English Corpus

Based on the values in Table 1 and Table 2, it can be drawn that the domain adaptability is not necessarily stronger when the corpus is larger. In English-Chinese bidirectional MT, LAW07 has the worst adaptability despite of its largest scale, while LAW09 has the best adaptability despite of its smallest scale.

3 Decremental Learning Framework

Inspired by the results of the statistical analysis of adaptabilities, we explored a decremental learning method which is a new idea of engineering level. Our decremental learning framework is shown in Figure 1. It mainly includes a domain filter, a quality filter and three identical NMT trainers. The preset data for implementing the framework includes common parallel sentences and domain text resources. Common parallel sentences refer to bilingual data that are large-scale, easily-available, domain-independent and errors-possibly-contained. Domain text resources can be monolingual or bilingual data, such as domain word sets, domain text documents, etc.

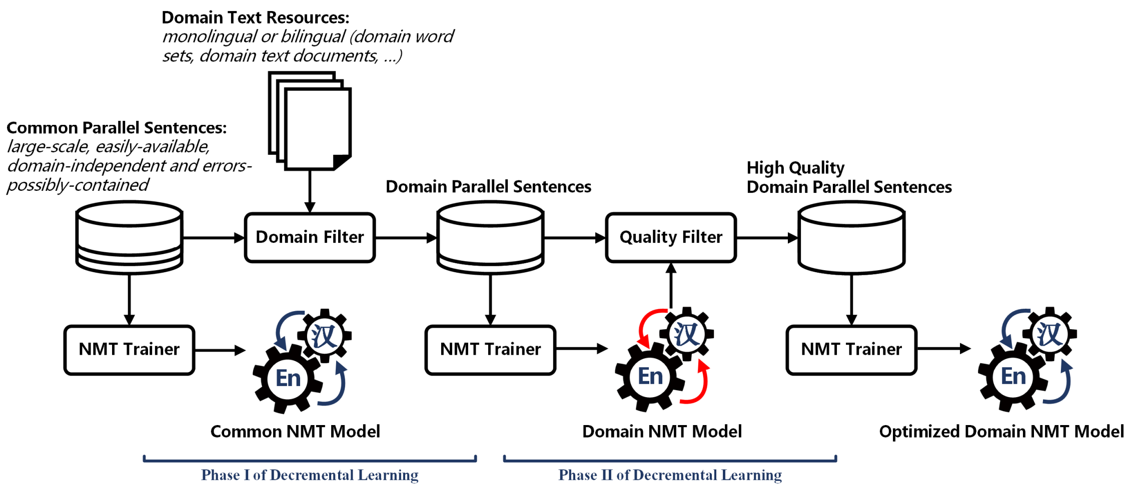


Figure 1: Decremental Learning Framework

This decremental learning framework is a meta structure independent of filtering algorithms, MT algorithms, and source languages and target languages. Here we take legal domain English-Chinese bidirectional MT as an example to implement the framework. Phase I classifies the sentence pairs in the common parallel sentences as legal domain and nonlegal domain. This is done by the domain filter with domain text resources. The result of this phase is parallel sentences. Then we train NMT models on an NMT trainer with common parallel sentences and domain parallel sentences, and respectively get an English-Chinese bidirectional NMT model and a legal-domain English-Chinese bidirectional NMT model. In Phase II we first translate every sentence pair of the domain parallel sentences with the legal-domain English-Chinese bidirectional NMT model, calculate the similarity between the original sentences and the sentences translated by the NMT model with Levenshtein string distance function in the quality filter, and delete the less similar sentence pairs based on the preset threshold. Hence we get high-quality domain parallel sentences and an optimized English-Chinese bidirectional NMT model in legal domain after training the model again. Three bidirectional English-Chinese NMT models (altogether six models) are trained in the whole framework, among which the common English-Chinese and Chinese-English models are only used for experimental comparison.

4 Domain Adaptation NMT Algorithm

To meet the actual domain adaptation NMT requirements, we designed a domain adaptation NMT algorithm based on our decremental learning framework. The algorithm is shown in Figure 2.

```
1.  // Domain Adaptation NMT Algorithm
2.  Input:    <String, String>[] train; // common parallel sentences of original training set
3.             <String, String>[] dev; // domain parallel sentences of development set
4.             <String, String>[] test; // domain parallel sentences of test set
5.             String[] dtr; // domain text resources
6.  Output:   Model odnmt; // optimized domain NMT model
7.  Model cnmt; // common NMT model only for comparison
8.  cnmt.st ← NMTTrainer.train(train, dev, test);
9.  cnmt.ts ← NMTTrainer.train(train, dev, test);
10. <String, String>[] train ← DomainFilter.filter(train, dtr); // domain parallel sentences
11. Model dnmt; // domain NMT model
12. dnmt.st ← NMTTrainer.train(train, dev, test);
13. dnmt.ts ← NMTTrainer.train(train, dev, test);
14. <String, String>[] mtout;
15. For Integer i ← 1 To train.size Do
16.            String TSen ← dnmt.st.translate(train[i].SSen);
17.            String SSen ← dnmt.ts.translate(train[i].TSen);
18.            mtout[i] ← <SSen, TSen>;
19. End For
20. <String, String>[] train ← QualityFilter.filter(train, mtout); // high quality domain parallel sentences
21. odnmt.st ← NMTTrainer.train(train, dev, test);
22. odnmt.ts ← NMTTrainer.train(train, dev, test);
23. Return odnmt.
```

Figure 2: Domain Adaptation NMT Algorithm

In Figure 2, **input** is the initial training set (train), development set (dev), test set (test) and domain text resource (dtr), and **output** is the opitimized domain NMT model (odnmt). The NMT training function (NMTTrainer.***train***) in lines 8, 9, 12, 13, 21 and 22 is implemented by the coder-decoder based on the attention mechanism. The input parameters are training set (train), development set (dev) and test set (test), and the output is an NMT model. The most critical two-phase filters are implemented by mature algorithms. The domain filter function (DomainFilter.***filter***) in line 10 is implemented by the text classification (SFITC) algorithm based on the String-Frequency Index (Liu et al., 2014). In Figure 2, the quality filter function (QualityFilter.***filter***) in line 20 is implemented with the algorithm of single-engine-based ensemble MT filtering (Liu and Wang, 2022).

5 Experiment

To verify the effectiveness and efficiency of decremental learning, we conducted an English-Chinese bidirectional NMT experiment in the legal domain.

5.1 Implement

We first implement the domain adaptation NMT algorithm according to the decremental learning framework. The parameters of the coder-decoder integrated in the algorithm mainly include the number of neurons (*num_units* = 512), the number of coder-decoder layers (*num_encoder_layers* = *num_decoder_layers* = 4), the number of training rounds (*epoch* = 10), the batch size (*batch_size* = 128) and the beam search width (*beam_width* = 10). The other parameters remain default values. Then we run the algorithm to obtain three English-Chinese NMT models and three Chinese-English NMT models. Finally, an interactive interface is added to form a web server as is show in Figure 3.

In terms of experimental data preparation, we first construct a parallel sentence bank in legal domain (PSB.ld) containing 100,000 pairs of English-Chinese parallel sentences and a bilingual word set in legal domain (BWS.ld) containing 76,792 items. According to the principle of simple random sampling, PSB.ld is divided into a development set (50,000 sentence pairs) and a test set (50,000 sentence pairs). BWS.ld is mainly used for the domain filter. Then we collect 21,942,400 pairs of English-Chinese parallel sentences (LAW07 corpus) from the Internet to form an initial training set.

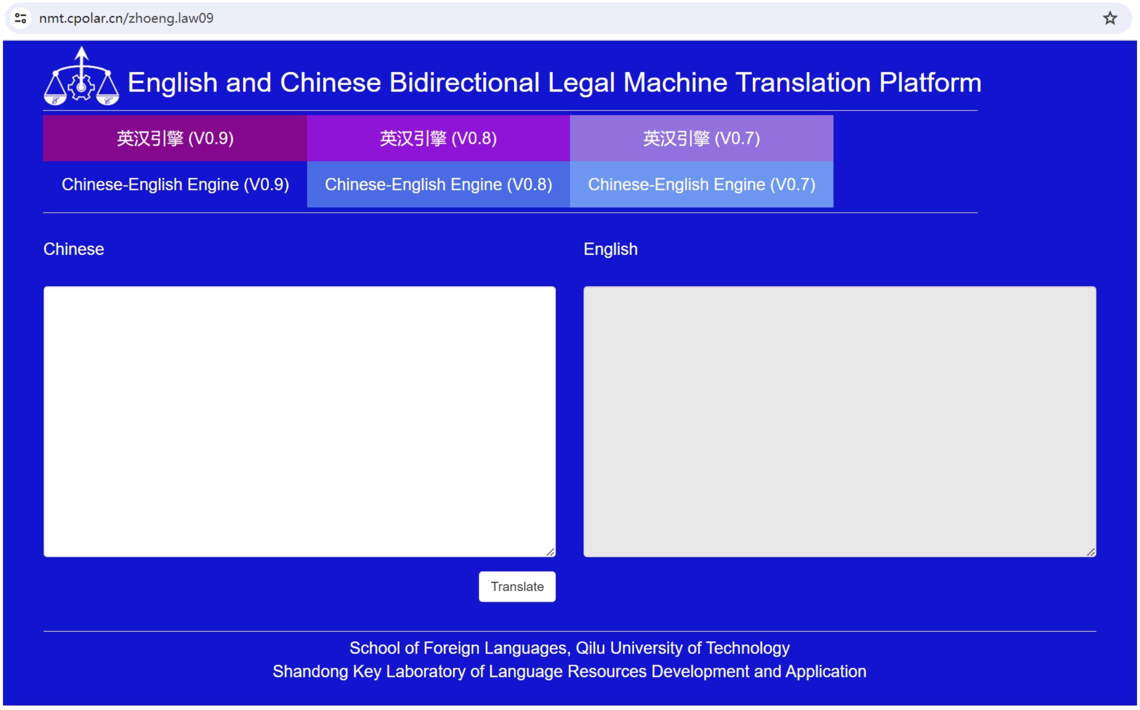


Figure 3: English and Chinese Bidirectional Legal Machine Translation Platform

In the course of the experiment, LAW07 successively produces domain parallel sentences (LAW08 corpus) and high-quality domain parallel sentences (LAW09 corpus) after decremental learning. We also reprocessed the data. For Chinese sentences, we separate each character with spaces. In the English-Chinese NMT experiment, we convert all English letters in lowercase and separate each word with spaces; while in the Chinese-English NMT experiment, we only separate each word with spaces without converting them.

5.2 English-Chinese NMT Experiment

The results of the English-Chinese NMT experiment are shown in Table 3. After phase I decremental learning, the 21,942,400 pairs of English-Chinese parallel sentences in the initial training set LAW07 corpus of unknown quality are reduced to 5,899,520 pairs in the domain training set LAW08 corpus. After phase II decremental learning, 5,710,080 pairs of English-Chinese parallel sentences of the high-quality domain training set LAW09 corpus are obtained. The BLEU values of the NMT models respectively trained with the three corpus are BLEU (LAW07) = 45.41, BLEU (LAW08) = 47.13, and BLEU (LAW09) = 47.88. It should be mentioned that the BLEU in this paper refers to the classic MT evaluation metric BLEU4. It can be seen that although the two-phase decremental learning reduces the scale of the training corpus, the MT performance does not reduce but improves. This also shows that decremental learning can effectively enhance the domain adaptability of the training corpus.

	number of sentence pairs of training set	**number of sentence pairs of development set**	**number of sentence pairs of test set**	**number of training steps**	**BLEU**
LAW07	21,942,400	50,000	50,000	1,714,250	45.41
LAW08	5,899,520	50,000	50,000	460,900	47.13
LAW09	5,710,080	50,000	50,000	446,100	**47.88**

Table 3: Results of English-Chinese NMT Model

The learning curves of the three NMT models during the training process in the English-Chinese NMT experiment is shown in Figure 4, with the abscissa being the training steps and the ordinate being BLEU values. The values on the ordinate of the end points of the three learning curves show that LAW07 model has the most training steps (1,714,250), and the number of the training steps of LAW08 model (460,900) is close to that of LAW09 model (446,100). On the other hand, the values on the abscissa of the end points of the three learning curves show that the performance of the LAW09 model is the best, which has the fewest training steps. This demonstrates that the result of decremental learning is a more efficient NMT model.

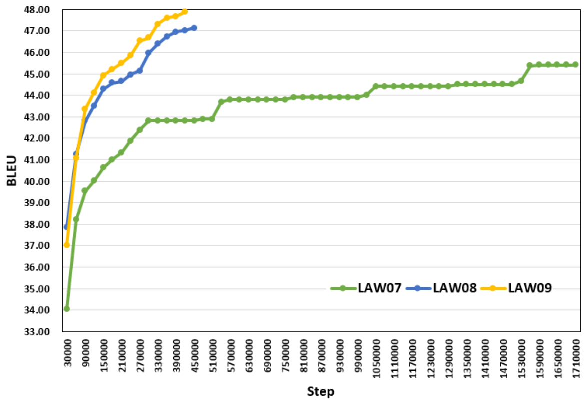


Figure 4: Learning Curves of English-Chinese NMT

5.3 Chinese-English NMT experiment

We also run the Chinese-English NMT experiment in the legal domain in reverse on the same corpus. Table 4 shows the results of the Chinese-English NMT experiment. Compared with Table 3, it is easy to know that the results of the Chinese-English NMT experiment are consistent with those of the English-Chinese NMT experiment. It is found that the translation performance of LAW09 (BLEU (LAW09) = 34.12) is better than that of LAW08 (BLEU (LAW08) = 33.81) and that of LAW07 (BLEU (LAW07) = 32.03). This demonstrates that decremental learning is also effective for Chinese-English NMT. We also find that the BLEU value of the Chinese-English translation is more than 10 points lower than that of the English-Chinese translation, both of which are trained with the same corpus. This is due to the need to maintain the uppercase and lowercase form of letters in Chinese-English NMT training, which increases data sparsity compared with all lowercase letter forms in English-Chinese NMT training.

	number of sentence pairs of training set	**number of sentence pairs of development set**	**number of sentence pairs of test set**	**number of training steps**	**BLEU**
LAW07	21,942,400	50,000	50,000	1,714,250	32.03
LAW08	5,899,520	50,000	50,000	460,900	33.81
LAW09	5,710,080	50,000	50,000	446,100	**34.12**

Table 4: Results of Chinese-English NMT Model

The learning curves of the three NMT models during the training process in the Chinese-English NMT experiment is shown in Figure 5. Compared to Figure 4, it is easy to find that the learning results of NMT are consistent in both Chinese-English and English-Chinese.

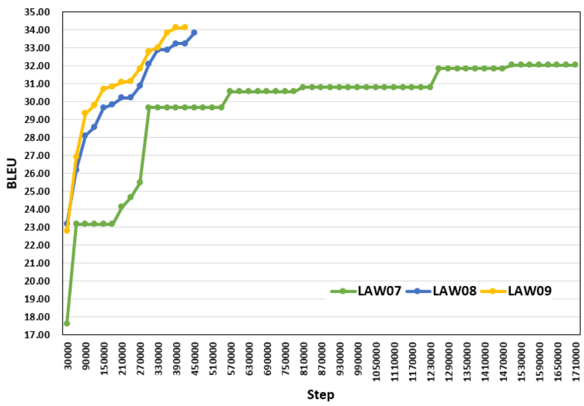


Figure 5: Learning Curves of Chinese-English NMT

5.4 Examples and Discussions

In order to show the effect of our English-Chinese NMT model more intuitively, we randomly selected three sentence pairs (Sen0, Sen1, and Sen2) for display. Table 5 lists the output translations of our three models LAW07, LAW08, and LAW09, as well as the output translations of Google, DeepL, and NiuTrans systems, where REF represents the human reference translation in the original sentence pair. A cursory glance shows that the output of all 6 models is acceptable. Careful discussion revealed that: “Pharmaceutical trading enterprise” in Sen0 is translated as “药品贸易企业” by DeepL, while the other 5 models are all translated as “药品经营企业”; “relevant judicial interpretations” in Sen1 is translated as “相关司法解释” by LAW07 and “有关司法解释” by DeepL, while other models add “的规定”; “regulating property relations” in Sen2 is translated as “规范产权关系” by LAW08, “调整财产关系” by the 3 general business models, and “规范财产关系” by our LAW07 and LAW09.

MT	Sen0.eng	Sen0.zho	Sen1.eng	Sen1.zho	Sen2.eng	Sen2.zho
LAW07	Pharmaceutical trading enterprise means an enterprise exclusively or partly engaged in the trade of pharmaceuticals.	药品经营企业是指专营或者部分从事药品贸易的企业。	The people's procuratorates shall, based on the facts, nature, circumstances and degree of harm to society of a crime, provide sentencing recommendations in accordance with the Criminal Law, Criminal Procedure Law and relevant judicial interpretations.	人民检察院应当根据犯罪的事实、性质、情节和危害程度，根据刑法、刑事诉讼法和相关司法解释，提出量刑建议。	The compilation of the civil code shall be advanced more quickly, the legal system relating to real rights, contract and property rights shall be improved, laws, regulations and provisions in contravention of fairness shall be reviewed, and equal protection shall be established as a basic principle of regulating property relations.	更快推进民法汇编，完善与权利、合同、财产权有关的法律制度，审查违反公平的法律、法规、规定，确立平等保护作为规范财产关系的基本原则。
LAW08		药品经营企业是指专营或者部分从事药品经营的企业。		人民检察院应当根据犯罪社会的事实、性质、情节和程度，依照刑法、刑事诉讼法和相关司法解释的规定提出量刑建议。		加快编纂《民法典》，完善与物权、合同、产权有关的法律制度，审查违反公平的法律、法规和规定，以平等保护为规范产权关系的基本原则。
LAW09		药品经营企业是指专营或者部分从事药品交易的企业。		人民检察院应当根据犯罪社会危害的事实、性质、情节和程度，依照刑法、刑事诉讼法和有关司法解释的规定提出量刑建议。		更快推进民法汇编，完善与物权、合同、产权相关的法律制度，审查违反公平的法律法规规定，建立平等保护作为规范财产关系的基本原则。
Google		药品经营企业，是指专门从事药品经营或者兼营药品经营的企业。		人民检察院应当根据犯罪的事实、性质、情节和社会危害程度，依照刑法、刑事诉讼法和有关司法解释的规定，提出量刑建议。		加快推进民法典编纂，完善物权、合同权、财产权法律制度，清理违背公平的法律、法规和规定，把平等保护确立为调整财产关系的基本原则。
DeepL		药品贸易企业是指专门或部分从事药品贸易的企业。		人民检察院应当根据犯罪的事实、性质、情节和对社会的危害程度，依照刑法、刑事诉讼法和有关司法解释提出量刑建议。		加快推进民法典编纂工作，完善物权、合同、财产权等相关法律制度，审查有违公平原则的法律法规和规定，将平等保护作为调整财产关系的基本原则。
NiuTrans		药品经营企业是指专门或者部分从事药品经营的企业。		人民检察院应当根据犯罪的事实、性质、情节和对社会的危害程度，依照刑法、刑事诉讼法和有关司法解释的规定提出量刑建议。		加快推进民法典编纂工作，完善物权、合同、财产权法律制度，对有违公平的法律法规和规定进行审查，把平等保护确立为调整财产关系的基本原则。
REF		药品经营企业是指经营药品的专营企业或者兼营企业。		人民检察院应当根据犯罪的事实、犯罪的性质，情节和对于社会的危害程度，依照刑法、刑事诉讼法以及相关司法解释的规定提出量刑建议。		加快推进民法典编纂工作，完善物权、合同、知识产权相关法律制度，清理有违公平的法律法规条款，将平等保护作为规范财产关系的基本原则。

Table 5: English-Chinese NMT Examples

Table 6 shows the effect of the Chinese-English NMT model. Comparing the reference translations, we find that the performance of each model has decreased, which is consistent with the BLEU evaluation results. In addition to the sparse data caused by keeping English case when we train NMT models, another potential reason is that the scale of English Token space is very different from that of Chinese Token space.

MT	Sen0.zho	Sen0.eng	Sen1.zho	Sen1.eng	Sen2.zho	Sen2.eng
LAW07	药品经营企业是指经营药品的专营企业或者兼营企业。	“ Pharmaceutical business enterprises ” means specialized enterprises that engage in drug management or concurrently operated enterprises.	人民检察院应当根据犯罪的事实、犯罪的性质，情节和对于社会的危害程度，依照刑法、刑事诉讼法以及相关司法解释的规定提出量刑建议。	The people's procuratorate shall, in accordance with the facts of the crime, the nature of the crime, the circumstances of the crime and the extent of harm to the society, submit sentencing suggestions in accordance with the provisions of the Criminal Law, the Criminal Procedure Law and the relevant judicial interpretations.	加快推进民法典编纂工作，完善物权、合同、知识产权相关法律制度，清理有违公平的法律法规条款，将平等保护作为规范财产关系的基本原则。	We shall accelerate the work of the codification of civil code, improve the legal system related to property rights, contracts and intellectual property rights, clarify the provisions of laws and regulations that violate the fairness of the law, and make equal protection the basic principle of regulating property relations.
LAW08		The term “ pharmaceutical trading enterprises ” refers to specialized enterprises or joint ventures that operate pharmaceuticals.		The people's procuratorate shall, according to the facts of the crime, the nature of the crime, the circumstances and the degree of harm to the society, put forward sentencing proposals in accordance with the Criminal Law, the Criminal Procedure Law and the relevant judicial interpretations.		To accelerate the promotion of the codification of civil codes, improve the legal systems relating to property rights, contracts, and intellectual property rights, clean up the provisions of laws and regulations that are in violation of fairness, and make equal protection a basic principle for regulating property relations.
LAW09		“ Pharmaceutical trading enterprises ” means specialized enterprises or joint ventures that engage in pharmaceutical products.		With regard to the facts of the crime, the nature of the crime, the circumstances and the degree of harm to society, the people's procuratorate shall put forward sentencing suggestions in accordance with the Criminal Law, the Criminal Procedure Law and the relevant judicial interpretations.		We shall accelerate the advancement of the compilation of civil law, improve the relevant legal systems relating to property rights, contracts and intellectual property rights, clean up the provisions of laws and regulations that are in violation of the law, and use equality protection as the basic principle for regulating property relations.
Google		A pharmaceutical business enterprise refers to an enterprise that specializes in or also operates pharmaceuticals.		The People's Procuratorate shall make sentencing recommendations based on the facts of the crime, the nature and circumstances of the crime and the degree of harm to society in accordance with the provisions of the Criminal Law, the Criminal Procedure Law and relevant judicial interpretations.		Accelerate the compilation of the Civil Code, improve the legal systems related to property rights, contracts, and intellectual property rights, clean up legal and regulatory provisions that violate fairness, and take equal protection as the basic principle for regulating property relations.
DeepL		A pharmaceutical business is a franchised or part-time business that deals in pharmaceuticals.		The people's procuratorate shall, on the basis of the facts of the crime, the nature of the crime, the circumstances and the degree of harm to society, make a recommendation on sentencing in accordance with the provisions of the Criminal Law, the Criminal Procedure Law and the relevant judicial interpretations.		Accelerating the codification of the Civil Code, improving the legal systems relating to property rights, contracts and intellectual property rights, clearing up legal and regulatory provisions that run counter to fairness, and making equal protection the basic principle governing property relations.
NiuTrans		A pharmaceutical trading enterprise refers to a franchised enterprise or concurrent enterprise that deals in pharmaceuticals.		The people's procuratorate shall, according to the facts, nature, circumstances and degree of harm to society of the crime, put forward sentencing suggestions in accordance with the provisions of the Criminal Law, the Criminal Procedure Law and relevant judicial interpretations.		Accelerate the compilation of the Civil Code, improve the legal systems related to property rights, contracts and intellectual property rights, clean up laws and regulations that violate fairness, and take equal protection as the basic principle for regulating property relations.
REF		Pharmaceutical trading enterprise means an enterprise exclusively or partly engaged in the trade of pharmaceuticals.		The people's procuratorates shall, based on the facts, nature, circumstances and degree of harm to society of a crime, provide sentencing recommendations in accordance with the Criminal Law, Criminal Procedure Law and relevant judicial interpretations.		The compilation of the civil code shall be advanced more quickly, the legal system relating to real rights, contract and property rights shall be improved, laws, regulations and provisions in contravention of fairness shall be reviewed, and equal protection shall be established as a basic principle of regulating property relations.

Table 6: Chinese-English NMT Examples

Based on the statistical results of the adaptabilities and the results of the English-Chinese bidirectional NMT experiment in the legal domain, the following findings can be drawn. (1) The BLEU value ranking of the English-Chinese bidirectional NMT model in the legal domain is fully consistent with the KLD and MMD numerical ordering of the corpora. This verifies that the KL divergence and maximum mean discrepancy can effectively measure the adaptability of the source domain corpus and the target domain corpus in adaptive NMT in a quantitative way. (2) Decremental learning is an effective domain adaptive strategy. In the process, the decremental learning of domain filtering can enhance the domain adaptability of the training corpus, and the decremental learning of quality filtering can improve the quality of domain-related translations of the training corpus. (3) The domain adaptation NMT algorithm only needs about a quarter of the original training steps and increase the BLEU values by over two points. This verifies that the decremental learning framework can efficiently optimize the performance of the NMT model. In summary, to train an NMT model for a specific domain, what matters are not the amount of the corpus but the high-quality domain-related corpus. Our decremental learning approach is a positive attempt for data engineering to convert readily available large-scale corpus of unknown quality to high-quality domain-related corpus.

6 Conclusion

This paper proposes a data centric approach focusing on the domain adaptation NMT problem, which gives full play to the large scale of parallel data of unknown quality and improves the domain translation quality through domain filtering and quality filtering. Finally, the effectiveness and efficiency of decremental learning is verified in the English-Chinese bidirectional NMT experiment in the legal domain.

Future research will mainly focus on domain knowledge modeling and NMT models based on domain knowledge intervention. We are going to construct an explicit multilingual domain knowledge graph, increase the neural computability of cross-language complex domain knowledge, and further improve the translation quality of domain adaptation NMT.

Acknowledgments

The research is supported by the Humanity and Social Science Research Project of Ministry of Education of China (No. 20YJAZH069, No. 20YJC740062), the New Liberal Arts Research and Reform Practice Project of Ministry of Education of China (No. 2021060049), the Postgraduate Education and Teaching Reform Research Project of Shandong (No. SDYJG21185), the Key Project of Undergraduate Teaching Reform Research of Shandong (No. Z2021323), the Key Project of the 14th Five-Year Plan of Educational Science of Shandong (No. 2023ZD021), and the 2024 Scientific Research and Innovation Platform Project of Xianda SISU (Language Data and AI Research, Xianda SISU).

References

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural Machine Translation: A Review of Methods, Resources, and Tools. *AI Open*, 1:5–21.

Chenhui Chu and Rui Wang. 2018. A Survey of Domain Adaptation for Neural Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319.

Tobias Domhan and Felix Hieber. 2017. Using Target-Side Monolingual Data for Neural Machine Translation through Multi-Task Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective Domain Mixing for Neural Machine Translation. In *Proceedings of the 2nd Conference on Machine Translation*, pages 118–126.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain Control for Neural Machine Translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 372–378.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost Weighting for Neural Machine Translation Domain Adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46.

Rui Wang, Masao Utiyama, Andrew Finch, Lemao Liu, Kehai Chen, Eiichiro Sumita. 2018. Sentence Selection and Weighting for Neural Machine Translation Domain Adaptation. *IEEE/ACM*

Transactions on Audio, Speech, and Language Processing, 26(10):1727–1741.

Adam Csaba Varga. 2017. Domain Adaptation for Multilingual Neural Machine Translation. Master Thesis, Saarlandes University.

Praveen Dakwale and Christof Monz. 2017. Fine-Tuning for Neural Machine Translation with Limited Degradation across in- and out-domain Data. In *Proceedings of the 16th Machine Translation Summit*, pages 156–169.

Zi-Yi Dou, Xinyi Wang, Junjie Hu, and Graham Neubig. 2019. Domain Differential Adaptation for Neural Machine Translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 59–69.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 385–391.

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization Techniques for Fine-Tuning in Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1489–1494.

Virginia Adams, Sandeep Subramanian, Mike Chrzanowski, Oleksii Hrinchuk, and Oleksii Kuchaiev. 2022. Finding the Right Recipe for Low Resource Domain Adaptation in Neural Machine Translation. arXiv preprint arXiv:2206.01137v1.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast Domain Adaptation for Neural Machine Translation. arXiv preprint arXiv:1612.06897.

Huda Khayrallah, Gaurav Kumar, Kevin Duh, Matt Post, and Philipp Koehn. 2017. Neural Lattice Search for Domain Adaptation in Machine Translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 20–25.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied Monolingual Data Improves Low-Resource Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Jiajun Zhang and Chengqing Zong. 2016. Bridging Neural Machine Translation and Bilingual Dictionaries. arXiv preprint arXiv:1610.07272v1.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-Supervised Learning for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1965–1974.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence Embedding for Neural Machine Translation Domain Adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 560–566.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic Data Selection for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410.

Danielle Saunders. 2022. Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey. arXiv preprint arXiv:2104.06951v2.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain Adaptation of Neural Machine Translation by Lexicon Induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001.

A. Tuan Nguyen, Toan Tran, Yarin Gal, Philip H. S. Torr, and Atılım Güneş Baydin. 2022. KL Guided Domain Adaptation. In *Proceedings of the 10th International Conference on Learning Representations*.

Wei Wang, Haojie Li, Zhengming Ding, and Zhihui Wang. 2020. Rethink Maximum Mean Discrepancy for Domain Adaptation. arXiv preprint arXiv:2007.00689v1.

Wuying Liu, Lin Wang, Mianzhu Yi, and Nan Xie. 2014. Active Multi-Field Learning for Spam Filtering. *Computing and Informatics*, 33(6):1400–1427.

Wuying Liu and Lin Wang. 2022. Ensemble Machine Translation to Filter Low Quality Corpus. In *Proceedings of the 26th International Conference on Asian Language Processing Proceedings*, pages 500–504.

應用大語言模型的提示詞工程於影音內容重點提取
(Application of Large Language Model-Based Prompt Engineering for Key Information Extraction from Audio-Visual Content)

Wei Ting Huang, Yu-Chen Liu and Ming-Hsiang Su

Department of Data Science, Soochow University, Taiwan
{bindy.huang, vincent93113, huntfox.su}@gmail.com

摘要

當今科技發達的時代，傳統傳播媒體已逐漸被網路影音所取代，而短影音更成為行銷、傳播資訊的重要媒介。爲協助影音創作者更有效率地將長影片製作成短影音以增加曝光度，本研究利用 BERT 模型探索大型語言模型對於篩選文本精華之能力。首先以人工抓取一百部影片字幕文本之精華，接著運用 GPT-4o 執行 4 種不同指令，蒐集 4 種不同的重要資訊篩選結果，最終將 4 種指令生成與人工篩選之文本進行相似度對比分析，歸納出最準確指令之特點。實驗結果發現給予大型語言模型較為明確地指令的確可以篩選出長影片的重要片段，使創作者能更有效率地將製作好的長影片剪輯為六十秒左右的短影音。

Abstract

In the era of advanced technology, traditional media has gradually been replaced by online video content, with short videos becoming a crucial medium for marketing and information dissemination. To assist content creators in efficiently transforming long videos into short clips to increase exposure, this study leverages the BERT deep learning model to explore the capability of large language models in extracting essential information from text. Initially, we manually curated key excerpts from the transcripts of over a hundred videos. Following this, we utilized GPT-4 to execute four distinct commands, collecting four different sets of important information. Finally, we performed a cosine similarity analysis between the texts generated by the four commands and the manually extracted excerpts, identifying the characteristics of the most accurate command. The experimental results revealed that providing large language models with more specific instructions enhances their ability to identify essential segments of long videos. This enables content creators to edit long-form videos more efficiently into short clips of approximately sixty seconds.

關鍵字：大型語言模型、BERT、短影音
Keywords: LLM, BERT, short video

1 緒論

在現今數位世代，網路已經深刻改變人們的生活習慣。根據 Lyndsi Stafford (2017)在富比士雜誌官網所述，人們觀看影片時能記住95%的訊息，而閱讀文字內容時只能記住 10%。由此可見，影片作為一種生動的資訊呈現方式，不僅在網路媒體中佔據著重要地位，也強化了資訊的傳遞效果。 自 2016 年 TikTok 上市後，短影音變得深受眾人喜愛。而 YouTube 和 Meta 為了因應短影音時代的趨勢，也在近兩年推出短影音服務。

短影音的流行使得觀眾的注意力變短，相較以往更沒耐心觀看完整長 影片。因此，近期許多創作者選擇在上傳長影片的同時，積極發佈短影音。透過短影音的快速傳播，吸引更多觀眾對長影片內容產生興趣，進而 點擊長影片，觀看完整內容。 然而，將長影片轉換為短影音仍然耗時費力。因此本研究將利用 GPT4o 模型挑選 YouTube 影片的重要片段，測試四種不同指令的挑選結果，並與人

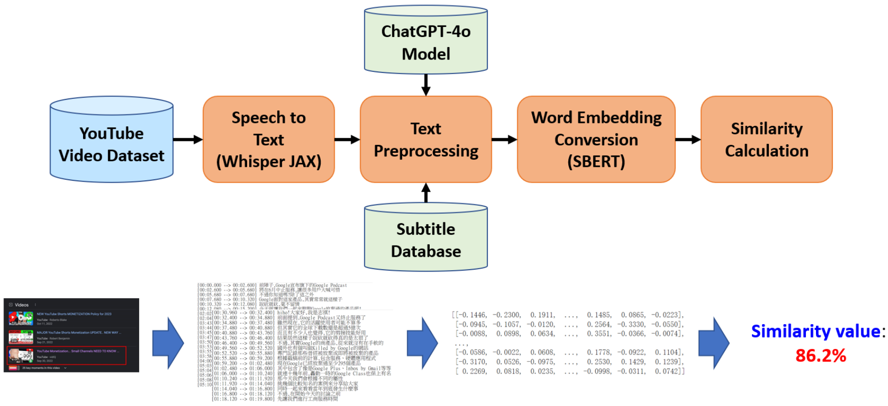


圖 1：研究方法流程圖

工選擇之片段進行比較，以檢測大型語言模型是否能自動剪輯出合適的短影音。

2 文獻回顧

2.1 短影音

短影音(short video)又稱為短影片、短視頻。目前在學術領域的定義仍有一些模糊之處，不過通常被定義為數十秒且豎頻形式呈現的影片。Fahao Chen 等人 (2023)認為短影音的影片時長通常小於一分鐘。另外，戴景麗 (2013)也將短影音定義為一分鐘內的短片。雖然目前對於短影音尚無明確統一的定義，不過考慮到本研究目的為吸引觀眾觀看完整長影片內容，因此本研究將製作目標之短影音定義為五十至六十秒的影片，以貼近實質觀看體驗。

2.2 大型語言模型(LLM)

由亞馬遜服務公司的《什麼是大型語言模型(LLM)？》一文可得知大型語言模型亦即 Large Language Model (LLM)，是一種大型的深度學習模型，可以用於生成式 AI，根據使用者提問回答相對應的內容。而 Csaba Veres (2022)指出大型神經語言模型在學習單字時是以連續向量表示機率函數，而非離散的詞彙，這有助於語言模型能學習更精確地語意表達模式，提供更正確的回應。由此推測，透過大型語言模型，應能自動判斷擷取，進而篩選出影片中的重點片段。

Raisa Islam 和 Owana Marzia Moushi (2024) 表示，GPT-4o 是目前 OpenAI 最新推出的模型，此模型相較於先前幾個版本的 ChatGPT 更精準，且效率也更高。且 GPT-4o 模型可以處理更廣泛知識的互動式問答，並引入了記憶功能，能記得使用者給予的特定細節。而 GPT-4o 模型也在 DROP 閱讀理解基準測試中拿到了 83.4%的分數，顯示出該模型應能判斷出長影片的重點片段，因此本研究使用 GPT-4o 作為篩選片段之語言模型。

2.3 BERT

Bidirectional Encoder Representations from Transformers (BERT) 是一種深度學習模型，根據 Koroteev M.V. (2021) 所述，BERT 模型在自動化文字處理展現了極高的精確度，該模型會使用向量的方式表示文本中的每個單字，以進行文字分類。該模型旨在對雙向文字表示進行深度初步學習，以便隨後在機器學習模型中使用。而相較於其他模型，BERT 更易於使用，只需在現有的神經架構中添加一個輸出層即可獲得在準確處理文本。

Aubrey Condor 等人(2021)表示儘管可以使用平均 BERT 輸出層從原始 BERT 模型導出句子向量，但此類方法會產生較差的結果。相較之下，Sentence-BERT (SBERT)所產生的句子向量在 SentEval 基準上優於其他最先進的方法，因此本研究選用以 BERT 為基礎設計的 SBERT 為文章轉向量之工具。

3 研究方法

本研究先將 YouTube 影片之語音內容轉換為文字檔，接著使用 Python 程式語言撰寫，利用 SBERT 模型將影片字幕檔之文字轉換為向量，再計算其餘弦相似度以比較不同提示詞(Prompt)之間的差異，過程如圖 1 所示。

本研究蒐集一百部 YouTube 之中文影片，透過 Whisper JAX (2024) 語音轉文字網站，將影

片中的語音轉換成文字。接著本研究以人工的方式進行資料預處理，將辨識錯誤的字詞更正，再自行建立字幕段落資料庫，用以分析大型語言模型篩選重要段落之能力。

接著使用 Python 程式語言撰寫程式碼，以 SBERT 模型將字幕文本轉換為文字向量。SBERT 是一個強大的語句嵌入工具，能夠將文本轉換成高維度向量，可用以比較不同文本之間的相似性或進行文本分類。最終本研究計算大型語言模型篩選出之重要文本段落與人工篩選之重要文本段落的餘弦相似度，以找出大型語言模型篩選重要段落之最佳提示詞樣板。

3.1 重要段落擷取

為分析大型語言模型篩選重要段落之能力，本研究除利用人工方式將字幕段落資料庫中之段落文本資料擷取出重要段落文本資料，此外，本研究亦將相同資料餵入(feed into) ChatGPT-4o 大語言模型中，利用本研究設計之四種提示詞用以篩選重要段落，如圖 2 所示。本研究使用之四種提示詞說明如下。

1. 你是一位專業的影片剪輯師。這份檔案是一部影片的字幕檔，請根據這份字幕檔的內容挑選出一段可以剪輯 60 秒短影音的段落。
2. 你是一位專業的影片剪輯師。這份檔案是一部影片的字幕檔，請根據這份字幕檔的內容挑選出 5 個段落，最終可以合在一起剪輯出 60 秒短影音的段落。
3. 你是一位專業的影片剪輯師。這份檔案是一部影片的字幕檔，請根據這份字幕檔的內容挑選出多個段落，最終可以合在一起剪輯出 60 秒（根據字幕檔的時間格式計算，相加後要介於 50~60 秒之間）短影音的段落。
4. 你是一位專業的影片剪輯師。這份檔案是一部影片的字幕檔，請根據這一整份字幕檔的內容挑選出多個段落。最終剪輯出 60 秒（根據字幕檔的時間格式計算，相加後要介於 50~60 秒之間）、包含影片重點的短影音，讓觀眾可以透過這支短影音就了解原本影片的重點。

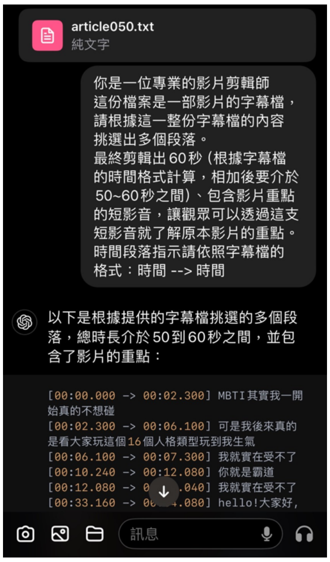


圖 2：ChatGPT-4o 對話示意圖

4 實驗結果與討論

4.1 資料集

本研究蒐集了一百部片長介於五到十分鐘之間的 YouTube 中文影片字幕檔，而影片主題類型包括電影解說、旅遊分享、美妝保養等各式內容。文本蒐集方式是利用以 OpernAI 所開發之 Whisper 製作的 Whisper JAX (2024) 語音轉文字網站讀取 YouTube 影片網址，使其自動將語音內容轉換為文字檔。為確保文字檔內容均與影片內容完全相符，透過網站將語音轉為文字後，再以人工方式進行核對，以確保文字檔與原始影片完全一致。

4.2 分析方式

本研究以 Python 3 環境撰寫程式碼，將重要段落文本資料透過 12 層 MiniLM v2 架構之 SBERT 模型轉換為向量。取得向量後再分析人工擷取之重要段落與 ChatGPT-4o 擷取之重要段落的餘弦相似度，並計算一百個文本的平均值與標準差，判斷四種不同提示詞所擷取之重要段落與人類篩選之重要段落最為相似。

實驗結果顯示，提示詞 4 的平均值約為 0.7583，是四者當中最高的，且標準差為 0.1825，相較其餘三種提示詞低，顯示提示詞 4 所擷取之重要段落最近似於人工擷取之重要段落。我們認為提示詞 4 相較其他三種指令給予的需求更為明確，包括秒數限制、必須出現影片重點，以及剪輯該短影音的目的。

表 1：不同提示詞與人工擷取重要段落相似詞評估

	Prompt 1	Prompt 2	Prompt 3	Prompt 4
平均值	0.6817	0.6786	0.7180	**0.7583**
標準差	0.2005	0.1931	0.1890	**0.1825**

5 結論與未來展望

在短影音時代的浪潮下，長影片創作者正面臨前所未有的挑戰，如何在有限時間內吸引觀眾注意力成為了核心問題。大型語言模型的生成技術為創作者提供了一個強大的工具，而如何有效利用這些工具則成為了一個關鍵課題。本研究探索了如何透過不同的指令來優化生成腳本的過程，以提升創作者的效率與作品品質。

我們的研究重點在於通過實驗驗證，不同的 prompt 如何影響影片文本的選取及生成短影音腳本的效果。我們首先以 4 種不同的指令來進行實驗，將影片文本上傳至 GPT-4o 模型，生成對應的 4 種短影音腳本。接著，我們以人工方式閱讀上百篇影片文本，挑選出適合節錄至短影音的片段，並使用 Bert 模型對生成的腳本進行相似度分析。通過對比平均數、標準差等數據，我們能夠找到最適合的指令，為創作者提供有力的參考依據。

本研究的未來規劃是為創作者提供一個系統化的方法，以最優化的指令來生成高質量的短影音腳本。我們期望透過這些實驗結果，能夠歸納出一套適用於各種影視及自媒體產業的公式模型，進一步推廣這些模型。再來，我們也計劃在未來研究自動剪輯的技術，使用人工智慧辨識及篩選長影片中之精華片段後，自動編排和剪輯。結合深度學習和情感分析技術，理解影片內容及觀眾偏好，更加靈活的創造出符合受眾的短影片，提高創作效率及市場競爭力。

References

Lyndsi Stafford. 2017. How to incorporate video into your social media strategy. Obtained from: https://www.forbes.com/sites/yec/2017/07/13/how-to-incorporatevideo-into-your-social-media-strategy. Access Date: 2024/07/15.

Fahao Chen, Peng Li, Deze Zeng and Song Guo. 2023. Edge-Assisted Short Video Sharing With Guaranteed Quality-of-Experience. *Journal of IEEE Transactions on Cloud Computing*, 11(1):13-24. https://doi.org/10.1109/TCC.2021.3067834.

戴景麗. 2013. *微視頻的內容定位與盈利模式*. PhD Thesis. 上海師範大學。

亞馬遜網路服務公司. 什麼是大型語言模型(LLM)?. Obtained from: https://aws.amazon.com/tw/what-is/large-language-model/. Access Date: 2024/07/15.

Csaba Veres. 2022. Large Language Models are Not Models of Natural Language: They are Corpus Models. *Journal of IEEE Access*. 10:61970-61979. https://doi.org/10.1109/ACCESS.2022.3182505.

Raisa Islam, Owana Marzia Moushi. 2024. GPT-4o: The Cutting-Edge Advancement in Multimodal LLM. *Authorea Preprints*.

Koroteev M.V. 2021. BERT: A Review of Applications in Natural Language Processing and Understanding. arXiv preprint arXiv:2103.11943. https://doi.org/10.48550/arXiv.2103.11943.

Aubrey Condor, Max Litster, Zachary Pardos. 2021. Automatic short answer grading with SBERT on out-of-sample questions. In *Proceedings of The 14th International Conference on Educational Data Mining*, pages 345-452. https://eric.ed.gov/?id=ED615495.

Whisper JAX: The Fastest Whisper API. Obtained from: https://huggingface.co/spaces/sanchit-gandhi/whisper-jax. Access date: 2024/07/15.

圖片描述技術輔助文本於可讀性模型準確度之研究

Evaluating the Accuracy of Image Captioning Technology in Enhancing Text Readability Models

Chieh-Hsuan Wu
Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology
Jessiewu.jjlin@gmail.com

Hou-Chiang Tseng
Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology
tsenghc@mail.ntust.edu.tw

Kuan-Yu Chen
Department of Computer Science and Information Engineering National Taiwan University of Science and Technology
kychen@mail.ntust.edu.tw

Berlin Chen
Department of Computer Science and Information Engineering, National Taiwan Normal University
berlin@ntnu.edu.tw

Yao-Ting Sung
Department of Educational Psychology and Counseling, National Taiwan Normal University
sungtc@ntnu.edu.tw

摘要

「適性閱讀」是將文本難度與讀者能力做適配，使讀者可以更容易理解文本內容。然而，隨著書籍的大量出版，由人工進行難易度評判將費時耗力，因而致力發展可讀性模型。過往文本可讀性模型的發展常以文字作為特徵，並未將圖片的內容作為評估文本難度的考量。然而，有其他領域指出，圖片能影響閱讀理解。受益此研究的啟發，本研究將利用圖片描述技術抽取出圖片的內容來做為特徵加入文本可讀性模型，以探討圖片內容的加入是否能夠提高可讀性模型的準確度。

Abstract

"Adaptive reading" is the process of adapting the difficulty level of a text to the ability of the reader so that the reader can understand the text more easily. However, with the large number of books being published, it is time-consuming and labor-intensive to manually assess the difficulty level, and therefore, efforts have been made to develop readability models. In the past, the development of text readability models was often characterized by text, and the content of pictures was not considered as a factor in assessing the difficulty of text. However, it has been suggested in other fields that pictures can affect reading comprehension. Inspired by this research, this study will utilize picture description techniques to extract the content of pictures as features to be added to the text readability model, in order to investigate whether the addition of picture

content can improve the accuracy of the readability model.

關鍵詞：文本可讀性、圖片描述、英文可讀性模型

Keywords: Text Readability, Image Captioning, English Readability Model

1 介紹

閱讀是學習重要的方式之一。閱讀不只能夠開闊視野，超越時空跟古人交談，獲取前人的智慧（曾志朗，2000）；亦能培養不同的能力，如閱讀過程中需要文字辨認與文意理解的能力，對於字詞與文句培養解碼能力；文意解讀方面則培養文本材料處理的能力（Carretti et al., 2020）。閱讀的好處不只是能力的培養，洪蘭（2005）於認知神經科學領域中，閱讀有助於刺激大腦神經的發展，預防大腦的退化；同時，透過閱讀，人們能夠減少因無知而帶來的恐懼，減輕挫折感（洪蘭，2005）；更是減輕壓力（Levine et al., 2022）。從文意理解方面至認知神經科學等，各界皆有學者從事相關研究，探討與閱讀有關之影響力或能力；可見閱讀對於個人的影響不可小覷。

然而，閱讀的重要性不僅對個人產生影響，更會影響到一個國家在國際競爭中的地位（OECD, 2009）。經濟合作暨發展組織（Organisation for Economic Cooperation and Development, OECD）於 2009 年的報告中指出：人力資本是當今最重要的資本形式，而國家的教育水準則是其經濟成長潛力的預測指標。透過閱讀行為的培養，不僅提升了個人的能力，也間接提高了國家的競爭力（Schleicher et al., 2009）。因此，各國開始重視閱讀能力的培育；並且透過參與國際間舉行的評量來了解各國學生閱讀表現與競爭力（張毓仁等，2011）。

國際學生能力評量計畫（Programme for International Student Assessment, PISA）針對參與國中的十五歲學生進行閱讀、數學和科學的評量；希望能為完成基礎教育學生的關鍵素養，提供具體的參考資訊（洪碧霞等，2018）。除此之外，另一項針對閱讀素養進行評量的促進國際閱讀素養研究（Progress in International Reading Literacy Study, PIRLS）為國際教育評估協會（The International Association for the Evaluation of Educational Achievement, IEA）所舉辦，為評估四年級學生的閱讀理解能力、閱讀素養成就及寫讀能力；並透過結果研究學生的能力及學習環境的影響（柯華葳等，2016）。由上述可知，兩個評量皆涵蓋閱讀素養的測驗及評估，可見閱讀於現今社會中的重要性；不僅如此，兩者皆藉由國際間評比的方式，提供各國關於教育成效的資訊，期盼能協助各國在制定與修改教育政策時有所依據（洪碧霞等，2018；柯華葳等，2016）。

我國對於閱讀素養的培育同樣重視。教育部於 2000 年開始推動閱讀教育，撰寫計畫並投入資源；從閱讀資源的普及，從都市至偏鄉的推廣，投入大量圖書資源與人力並且設立圖書館等適合閱讀的環境；至「提升國民中小學學生閱讀教育實施計畫」的三年期計畫（教育部，2015）。此計畫因應「十二年國民基本教育課程綱要」之理念為提升國中小學生閱讀素養（教育部，2018），除了將閱讀資源整合共享、促進學生們參與多元的閱讀行動外；更是推動「自主閱讀適性學習」，當學生透過數位網路取得多元文本或大量訊息的同時，培養資訊素養與辨別能力（教育部，2015）。

為回應適性閱讀的議題，各級學校紛紛推行相關競賽及活動，邀請學生與家長一同加入行列。然而，文本的挑選適合於否，成為重要的議題。因為，面對文本與書籍等大量出版，靠著專家以人工的方式評判文本難易度不僅需要花費大量時間，更是費力且須以龐大開銷支持；

為解決此現象，文本可讀性模型的開發，使得人們可藉著自動化工具大量檢測文本難易度，來處理如此巨量的資料。事實上，文本可讀性的相關研究早在 20 世紀初，便有學者著手相關研究。如 Thorndike（1921）將整理英文詞彙之使用頻率，並將其作為頻率表，欲提供做為文本難度的標準；Lively 與 Pressey（1923）則依據該頻率表發展出第一個可讀性公式；Flesch Reading Ease（Flesch, 1948）將句長以及詞彙音節數納入做為特徵，等等選擇淺層的語言特徵發展可讀性公式的傳統可讀性模型（翁詩謠 等，2019）；除此之外，亦有使用各種語言特徵與機器學習模型相結合，這些文本被轉換成許多語言特徵，這些特徵被輸入到 Support Vector Machine（SVM）和 Multilayer Perceptron（MLP）等簡單模型中（Schwarm & Ostendorf, 2005; Vajjala & Meurers, 2012）；直到現今仍有許多學者持續研究相關領域，可見可讀性模型議題十分重要（Patel et al., 2023; Murgia et al., 2023）。

由上述可知，目前可讀性模型大多著眼於文字、字義的研究（Collins-Thompson, 2014），對於探討圖片資訊對於可讀性的影響尚未有相關研究；然而，在其他領域上已有許多研究證實視覺與閱讀理解之間互有影響力。例如：科學圖文閱讀領域的研究指出，於科學文本中添加圖像會影響讀者的閱讀行為，但無法保證學習表現良好；透過教學介入與適當引導，有助於閱讀理解之提升（王孜甯、簡郁芩，2022）；當文字將與之相呼應的圖片作為提示，或許可以減少認知負荷並促進資訊整合。當語言與圖像特徵連結於同一材料時，根據多媒體認知理論，自文字與圖片中學習成果將比只從文字中學習更佳（Abdulrahaman et al., 2020）。由上述研究可知文本中的圖片會影響讀者的理解與閱讀，透過圖片輔助作為提示幫助讀者在文本上的解讀，或許能提高讀者對於文本的理解。本研究受益於上述研究的啟發，在可讀性模型中加入圖片內容作為特徵，以探討是否能提升可讀性模型的準確率；換句話說，本研究僅探討加入圖片內容對於可讀性模型效能的影響，並非探討將圖片內容的資訊融入原始文章中對於讀者的影響。本研究將使用圖片描述（image captioning）技術，從圖片生成語意，作為可讀性模型特徵，並透過回歸模型探討圖片敘述是否能提高模型準確度；期盼證實圖片內容做為特徵能夠提高可讀性準確率。本論文接下來的安排如下：第二章將回顧與文本可讀性模型以及圖片擷取技術相關之文獻；第三章呈現實驗材料與流程相關內容與操作方式；第四章則依實驗之結果進行分析與討論；最後第五章將提出未來可研究之方向。

2 文獻回顧

可讀性（readability）是指讀者能夠理解文本的程度（宋曜廷等，2013）。至於，可讀性的高低則受文字特徵以及讀者本身差異而有所影響（Collins-Thompson, 2014）。由於讀者的能力和背景各不相同，對於文本的理解也會有差異。因此，不同領域常使用可讀性模型來評估文本的難易程度，如 Flesch-Kincaid Grade Level（FKGL）大多用於教育、醫學領域，Lexile 或 Advantage-TASA Open Standard（ATOS）等則用於商業領域（CLEAR, 2023）；除以上模型以字詞出現頻率作為可讀性模型特徵外；Mesmer（2020）表示，Degrees of Reading Power、Reading Maturity Measure（RMM）和 Source Rater 以及上述提及之 Lexile 與 ATOS，為五種常用的可讀性公式；同樣採用「句長」、「音節數量」、「字詞出現頻率」等與文字、語意相關作為特徵。

可讀性特徵的研發十分多元，自字數等文字表面特徵至語意、上下文句等較深入意義的特徵；密西根大學教授 Collins-Thompson（2014）將可讀性特徵進行回顧與整理，提出特徵類別

圖，將特徵以低到高分為六個層次：文本易讀性、詞彙/語意、文法、文本結構、高級語意學、讀者興趣與背景；可見大多以文字本身及語意作為特徵；然而，除了文字上的特徵外；亦有其他的研究從排版（layout）的角度出發，如「字體」、「格式」、「間距」等（Collins-Thompson, 2014），探討視覺上對於可讀性的影響。

由上可知，過去可讀性模型大多使用文字作為特徵，雖然有部分研究使用視覺相關的特徵，而這些研究也只將圖片排版納入考量（Collins-Thompson, 2014）。然而，其他領域研究指出圖片內容能輔助讀者的閱讀理解，如王孜甯與簡郁芩（2022）證實將圖片加入科學文本中，透過合適的引導幫助讀者增加對圖文理解程度（王孜甯，簡郁芩，2014）；張菀真與辜玉旻（2011）則透過放聲思考法及附圖的自然文本，探討國小學生閱讀策略及插圖運用狀況，發現高閱讀理解能力者能整合圖文訊息且運用插圖提高理解程度，由此可知，圖片內容能幫助閱讀理解。受上述研究的啟發，本研究欲使用圖片描述技術抽取圖片內容，並使用圖片內容做為可讀性模型特徵，了解圖片是否能作為文本輔助，提高可讀性模型準確率。

欲加入圖片內容作為特徵，需事先將圖片內容轉譯成文字。資訊領域中，圖片描述技術可從圖片中擷取特徵，進行圖片辨認，過去做的研究大多是關於符號與文字辨識以及圖片分類（Wang et al., 2023; Wang et al., 2023）。然而，近年來多模態大型語言模型（Multimodal Large Language Model，MLLM）的發展，能夠從輸入多種媒體型態，如圖片、音訊、影片，輸出相對應的圖片或文字（Huang et al., 2024）；本研究將利用圖片描述技術，擷取圖片內容作為可讀性特徵，期能夠提升可讀性模型效能。

3 實驗

本研究欲使用圖片描述技術，從圖片中擷取圖片內容成文字，將之視為特徵加入至可讀性模型中，期望可以透過圖片內容的加入提升可讀性模型之準確率。

3.1 實驗資料

本實驗使用由 Crossley 等人（2023）提出 CommonLit Ease of Readability（CLEAR, 2023）資料庫作為實驗材料；此資料庫請專業教師們為 4724 筆資料挑選出合適資料文本之摘要，並透過 Bradley-Terry 模型取得難度；內容依據符合英文母語者三年級至十二年級課程使用為準則做資料收集，摘錄 250 年來兩種不同寫作流派，供更多領域的學者作使用（Crossley et al., 2023）。並將文本主要分為資訊文本以及文學文本，且將資訊文本細分為：歷史/社會科學、科技以及技術學科等。隨著文本種類的豐富，此資料庫亦收入不同難易程度的文本；本實驗以 Lexile reading band（Smith, 1989）做為資料篩選之標準，Lexile 難度與學習者年齡對照表如表 1。

年級	學期末	
	50%	90%
幼稚園	BR160L	150L
1 年級	165L	570L
2 年級	425L	795L
3 年級	645L	985L
4 年級	850L	1160L
5 年級	950L	1260L
6 年級	1030L	1340L
7 年級	1095L	1410L
8 年級	1155L	1470L
9 年級	1205L	1520L
10 年級	1250L	1570L
11 年級	1295L	1610L
12 年級	1295L	1610L

表 1. Lexile 難度與學習者年齡對照表

以 Lexile 網頁預設的建議，其中記載以英文為母語學習者於學期初及學期末各百分比學生

之閱讀能力；隨著年紀增加，學生能接受的文本難度也逐漸提高，這與學生的識字程度密切相關。因此，在識字量較少的低年級，文本通常會配有圖片，以幫助學生更好地理解內容。

在此，本研究選取閱讀資料中圖片較多之年紀：英文母語學習者一年級至四年級，選取 Lexile 300~900 的資料；於其中挑選閱讀材料含有圖片且圖片內容符合文本文意之資料，共 155 筆資料，作為本次實驗之測試資料；而剩餘未符合上述條件之 4569 筆資料作為訓練資料。

3.2 實驗工具

3.2.1 可讀性特徵：Global Vectors

本研究使用史丹佛大學提出的 Global Vectors 語言模型做為訓練可讀性模型的特徵（Pennington et al., 2014）。該模型將單字表示為連續空間中的向量，並以 300 維度作為表示；基於全域對數雙線性回歸（global log-bilinear regression）與統計特性，計算成本與其他模型相比較為低且易執行；因此，有許多研究使用此語言模型或進行優化（Kulkarni et al., 2020; Sakketou & Ampazis, 2020）。本研究使用此語言模型抽取可讀性特徵，將文本中的詞彙，以查表方式找出詞彙的語意向量，並將所有詞彙的向量以累加的方式得出一個 300 維度的文章語意向量來做為特徵。詳細解釋於 3.3 實驗流程當中加以解釋。

3.2.2 視覺-語言模型：ChatGPT-4

本研究使用 OpenAI 開發的 ChatGPT 作為圖片描述生成工具，此模型能根據使用者輸入的內容，利用人工智慧技術自動生成敘述，以回答使用者提出的各種問題，介面採用對話的方式，模擬與真人對話的情況，使情境更為擬真（Open AI, 2023； Kalla et al., 2023）。過去，ChatGPT 只接受文字輸入，相關研究大多以使用者體驗與 ChatGPT 所帶來的影響作為探討（Fuchs, 2023; Kalla et al., 2023）然而，GPT-4 模型的誕生，打破了只能輸入文字的限制。GPT-4 是一個多模態的大型語言模型，基於 Transformer 模型進行訓練，能夠處理圖片及文字的輸入，並生成文字敘述；因此，為研究圖片內容對於讀者理解的影響，本實驗使用了 ChatGPT 中的 GPT-4 模型（OpenAI, 2023）來為圖片生成描述。

3.2.3 可讀性模型：Decision Tree Regressor

為更精準地預測文本的可讀性結果，本研究選用 Scikit-learn 函式庫（Pedregosa et al., 2011）中的 Decision Tree Regressor 決策樹回歸函數進行模型建置。該函數具有減少記憶體占用、提升計算效率和速度的特性，並且可以通過調整權重來進行不同的實驗，適用於許多不同產業的預測（Rahul et al., 2013； Joshi et al., 2020）。

3.3 實驗流程

本實驗使用 CommonLit Ease of Readability（CLEAR, 2023）資料庫作為實驗材料，以 Lexile（Smith, 1989）做為文本難易度指標選取 Lexile 300～900 之資料，從中選取閱讀材料中涵蓋圖片且圖片內容與文本同源之材料，作為本實驗的測試資料；其餘則作為訓練資料。本實驗將基於圖片影響閱讀理解，透過圖片特徵的有無作為變因，將測試資料分為含有文本及圖片敘述的實驗組以及單純文本的對照組。對照組實驗流程如下圖 1 所示。

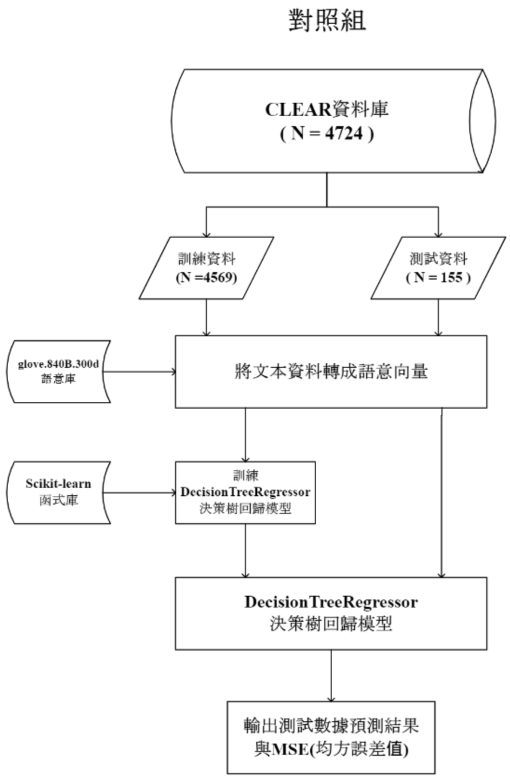


圖 1. 對照組流程圖

將文本測試資料以及訓練資料分別透過程式進行文字轉語意向量的轉換；引用 Global Vectors for Word Representation 概念模型（Pennington et al., 2014）中的 glove.840B.300 語言模型，運作過程如圖 2 所示。

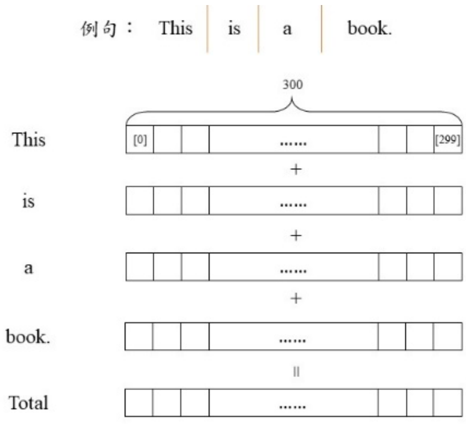


圖 2. 文本轉譯語意向量步驟

每筆資料依空格切分為單詞，逐一讀取每個字詞，至語言模型中搜尋該字詞 300 維度的語意向量並回傳，加到計算文本 300 維度語意向量的矩陣容器中；累加該文本中每一單詞的 300 維度語意向量，直到讀完整篇文章，即可獲得該筆資料總共的語意向量，作為訓練可讀性模型的特徵。

將語意向量轉換完成的資料，加入 CommonLit Ease of Readability（CLEAR, 2023）中提供的 Bradley-Terry 值作為正確答案，並將訓練資料輸入以 Decision Tree Regressor（Pedregosa et al., 2011）建立的決策樹回歸模型，進行可讀性模型的建置；最後，將同樣完成語意向量轉換的對照組資料輸入模型進行可讀性預測；即可得到各文本的預測值以及整體與正解的均方根誤差（Mean Squared Error, MSE）。

原文本加入圖片內容轉述的實驗組流程，與對照組大致相同，實驗流程如圖 3 所示。

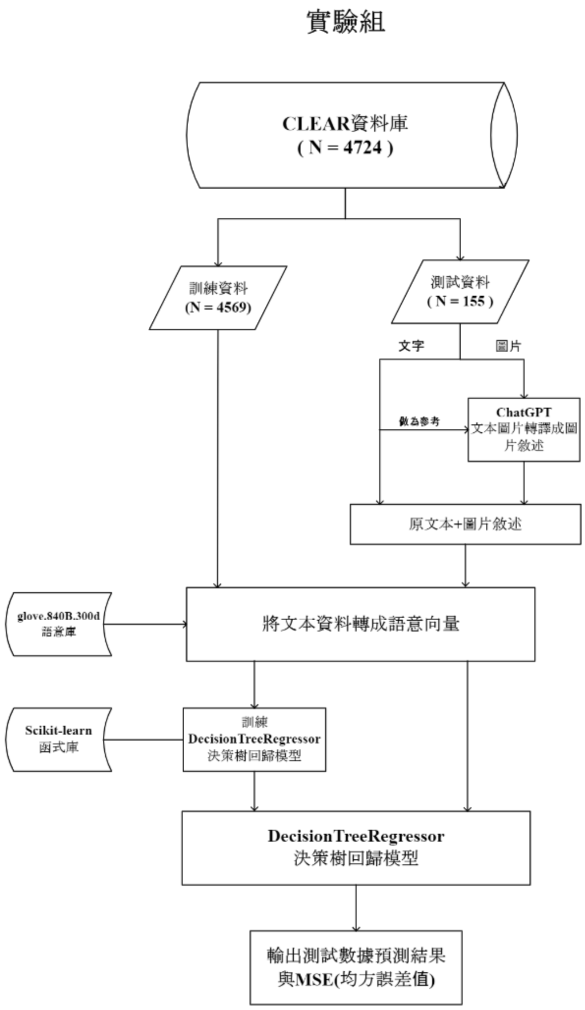


圖 3. 實驗組流程圖

本研究根據圖片敘述的有無作為實驗與對照組的變因，因此實驗組使用 ChatGPT4.0（Open AI, 2023）的視覺—語言模型，將原文做為參考，同時輸入圖片取得圖片描述，並將輸出的描述文字合併至原文本最後的位置，做為特徵的增加；其他流程則與對照組相同：使用 glove.840B.300 語言模型將文本轉換為語意向量；並使用與對照組相同回歸模型，輸入測試資料進行預測。將對照組與實驗組結果相比，判斷是否實驗組有低於對照組，驗證當圖片文意加入時是否可以更靠近真實文的文本難度。

4 實驗結果與未來展望

4.1 實驗結果

為證實圖片的內容特徵能提升可讀性模型的效能，本研究將圖片內容透過圖片敘述生成技術，使用視覺-語言模型將圖片內容依照文本生成補述，作為可讀性模型特徵，以提高可讀性模型準確率，實驗結果由表 2 可知。

	MSE 值
實驗組（文本加入圖片描述）	1.60
對照組（純文本）	2.59

表 2. 實驗結果（N＝155）

實驗組所預測出與正解之均方誤差為 1.6，對照組則為 2.59，實驗組之均方誤差低於對照組，兩組數值相差 0.99，表明本研究將視覺-語言模型所生成的圖片敘述加入文本的做法，確實能夠提升可讀性模型在文本難度評估上的表現，圖片內容對於提高文本可讀性的貢獻顯著。除此之外，CommonLit Ease of Readability（CLEAR, 2023）語料庫包含各式學科，由於各學科類型不同，圖片的表徵結構與內容及使用方式也有所差異，因此文本與圖片內容的圖文相關程度也是本篇的探討方向。本研究按語料庫內提供各資料之學科，分類成：History（歷史）、Science（科學）、Technology（科技）、Bio（生物）以及空格；其中，本研究所使用的未註明學科資料大多來自於 gutenberg.org 網站，內容以黑白圖文小說居多，其次則來自其他兒童圖書網頁，內容則以文學與科學故事為多。各學科於實驗組與對照組表現如表 3 所示。

學科	篇	實驗組 MSE	對照組 MSE
未註明學科	68	1.57	2.88
History	22	1.47	3.91
Science	47	1.56	2.3
Technology	21	1.95	0.88
Bio	1	1.15	2.25

表 3. 各項學科與各組 MSE 比較（N=155）

由表知，歷史、科技、生物及未註明學科，MSE 值皆呈現實驗組低於對照組，意謂著在上述學科中，加入圖片內容作為特徵，確實能夠幫助可讀性模型提高準確度；然而，反觀文本類型為科學者，其實驗組之 MSE 值則低於對照組，表現不太理想。

4.2 討論

圖片內容敘述作為特徵加入文本末端，是否有助於提高可讀性模型準確度，可以從整體結果的 MSE 值比較得到證實，擷取圖片內容成為特徵，確實降低模型誤差使預測更精準。表 3 將結果分成各科，由表 3 可知，不同學科會有不同的影響；多數科目除科技外，以 MSE 角度觀察，實驗組皆低於對照組，其中以歷史及未註明學科兩項，實驗組與對照組 MSE 值落差最大、表現最佳。歷史學科多以歷史名畫、偉人雕像居多，其文本資料大多來自維基百科；其內容於網路上有大量的資料，圖片描述過程中，有較多的內容可供參考與生成；未註明學科大多以黑白手繪圖畫居多，來自 gutenberg.org 網站；資料內容大多為 Lexile500，適合英文母語者小學一至二年級程度，內容較為簡單且圖片複雜度低，在圖片描述上有較佳的轉譯結果。

反觀本次表現最差的學科—科技，其資料雖然大多來自維基百科且包含真實照片，但因其

內容包含物品實體圖、相關原理流程圖及示意圖；當讀者在進行閱讀時，圖片的作用大多是使用圖式輔助讀者理解運作過程及使用方式；以資料 Cathode Ray Tube（陰極射線管）為例，表 4 為該筆資料的圖片及圖片描述。

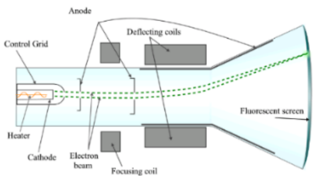
序號：1937	Lexile: 900	主題：Cathode ray tube
圖片		
		
圖片描述		
This diagram illustrates the internal structure of a CRT, showing the electron gun and path to the phosphor screen where the image is displayed.		

表 4. 科技學科其一資料圖片以及圖片敘述

此資料文本內容詳細描述陰極射線管的背景及運作方式，然而由表四可知圖片描述唯有指出此圖為陰極射線管，並且以較精簡的文句敘述過程。由此可知，兩者內容雖然相同；但在用詞上，圖片描述卻使用原文並沒有使用的字眼，如:「路徑」(path)、「內部結構」(internal structure）等，因而產生差異；另外，圖片描述內容十分簡潔，無法真實模擬讀者在閱讀時，對於整個運作流程理解的過程，使圖片描述無法成為文本輔助。

4.3 結論與未來研究

在與閱讀相關的研究中指出，在閱讀資料中圖片能夠影響讀者理解，輔助讀者理解文本內容。然而，過去的可讀性模型所使用的特徵卻大多停留在文字的層次，圖片的資訊鮮少使用。因此，本研究使用圖片描述技術，將圖片內容轉譯並加入原始文本中，進行文本難易度預測。實驗結果顯示額外加入圖片內容敘述的實驗組，MSE 誤差低於純文本的對照組；證實圖片內容能輔助讀者閱讀，提升讀者對於文本的理解。另一方面，使用圖片內容作為資料特徵，並且加入原有的特徵當中，不只提高可讀性模型準確率，使可讀性模型能更接近專家評判，增加可信度；也因本研究單純將圖片特徵加入至測試資料當中，可讀性模型準確率立即提高，得知不須更改或重新訓練可讀性模型內容，亦可提到準確率；由此可知，本研究所研發之技術有潛力直接應用於其他可讀性模型，也許同樣能夠提高準確率。換言之，無須使用本研究所建立的模型，只需於欲使用的模型中，將測試資料加入新特徵，或許亦可完成任務，使文本難易度的測量更加省時省力。

然而，圖片內容輔助文本理解在各科上影響力略有不同；此現象造成的原因或許牽涉本研究資料量不足，僅能以少量數據做為討論；各學科資料不均，導致無法更深入了解各學科之間圖片輔助文本的影響力，以及文本型態和圖片風格差異所帶來解讀上的影響。

有鑑於此，為更加客觀的了解圖片輔助文本在各學科上，幫助讀者理解的程度，增加資料筆數，抑或提升資料豐富度如：增加學科種類、分析各學科圖片資訊等，是接下來進一部探討的方向；除此之外，不同語料庫及語言模型的選擇，查驗語料庫或語言模型的不同是否影響圖片輔助文本於模型的預測結果亦是未來可研究的方向。期盼能更進一步了解圖片內容做為文本輔助幫助讀者理解與作為新特徵提高可讀性模型準確度相關的內容。

致謝

This work was supported by the National Science and Technology Council of Taiwan under Grants NSTC 113-2410-H-011 -001, NSTC 112-2628-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.

參考資料

王孜甯、簡郁芩（2022）。科學圖文閱讀眼動研究之系統回顧. 國立臺灣師範大學教育心理與輔導學系 教育心理學報，53（4），773–800. https://doi.org/10.6251/BEP.202206_53（4）.0001

日本文省部（2007）。日本文省部科學省. https://www.mext.go.jp/

柯華葳、張郁雯、劉子鍵、詹益綾、陳家興、賴明欣（2016）。台灣國際閱讀素養調查--PIRLS 2016（第 4 年）

曾志朗（2000）。閱讀是多元智慧成功的基本條件。教師天地，106，4-5。

洪蘭、曾志朗（2000）兒童閱讀的理念— 認知神經心理學的觀點. 現代教育論壇— 兒童閱讀的理念與策略, 國立教育資料館, 台北市立師範學院, 2000-11-15, 2000.

張毓仁、柯華葳、邱皓政、歐宗霖、溫福星.（2011）. 教師閱讀教學行為與學生閱讀態度和閱讀能力自我評價對於閱讀成就之跨層次影響：以 PIRLS 2006 為例. 教育科學研究期刊, 56(2), 69-105.

臺灣 PISA 國家研究中心主編（2018）.PISA2018 台灣學生表現.心理出版社.

Abdulrahaman, M. D., Faruk, N., Oloyede, A. A., Surajudeen-Bakinde, N. T., Olawoyin, L. A., Mejabi, O. V., ... & Azeez, A. L. （2020）. Multimedia tools in the teaching and learning processes: A systematic review. Heliyon, 6（11）. https://doi.org/10.1016/j.heliyon.2020.e05312

Anne E. Mesmer, H., Hiebert, E. H., Cunningham, J. W., & Kapania, M. （2020）. Does one size fit all? Exploring the contribution of text features, text content, and grade of use on comprehension. Reading Psychology, 42（1）, 42-72. https://doi.org/10.1080/02702711.2020.1861864

B. A. Lively & S. L. Pressey, "A method for measuring the vocabulary burden of textbooks," Educational administration and supervision, vol. 9, pp. 389–398, （1923）

Collins-Thompson, K. （2014）. Computational assessment of text readability: A survey of current and future research. ITL-International Journal of Applied Linguistics, 165（2）, 97-135. https://doi.org/10.1075/itl.165.2.01col

Crossley, S., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. （2023）. A large-scaled corpus for assessing text readability. Behavior Research Methods, 55（2）, 491-507.

Carretti, B., Toffalini, E., saponaro, C., Viola, F., & Cornoldi, C. （2020）. Text Reading Speed in a Language with a Shallow Orthography Benefits Less from Comprehension as Reading Ability Matures.British Journal of Educational Psychology,90,91–104. https://doi.org/10.1111/bjep.12307

Fuchs, K. （2023, May）. Exploring the opportunities and challenges of NLP models in higher education: is Chat GPT a blessing or a curse?. In Frontiers in Education （Vol. 8, p. 1166682）. Frontiers Media SA. https://doi.org/10.3389/feduc.2023.1166682

Gupta, A., Bansal, A., & Roy, K. （2021, May）. Solar energy prediction using decision tree regressor. In 2021 5th International Conference on Intelligent Computing and Control Systems （ICICCS）, 489-495. IEEE.

Huang, D., Yan, C., Li, Q., & Peng, X. （2024）. From Large Language Models to Large Multimodal

Models: A Literature Review. Applied Sciences, 14 （12）, 5068. https://doi.org/10.3390/app14125068

Joshi, N., Singh, G., Kumar, S., Jain, R., & Nagrath, P. （2020）. Airline prices analysis and prediction using decision tree regressor. In Data Science and Analytics: 5th International Conference on Recent Developments in Science, Engineering and Technology, REDSET 2019, Gurugram, India, November 15–16, 2019, Revised Selected Papers, Part I 5 , 70-186. Springer Singapore.

Kincaid, J. P., Fishburne, L. R. P., Rogers, R. L., & Chissom, B. S. （1975）. Derivation of new readability formulas （automated readability index, Fog Count and Flesch Reading Ease Formula） for navy enlisted personnel. Millington, TN: Navy Research Branch.

Kulkarni, S., Katariya, J. K., & Potika, K. （2020, December）. Glovenor: Glove for node representations with second order random walks. In 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining （ASONAM）, 536-543. IEEE.

Kalla, D., Smith, N., Samaah, F., & Kuraku, S. （2023）. Study and analysis of chat GPT and its impact on different fields of study. International journal of innovative science and research technology, 8 （3）.

Levine, S. L., Cherrier, S., Holding, A. C., & Koestner, R. （2022）. For the love of reading: Recreational reading reduces psychological distress in college students and autonomous motivation is the key. Journal of American College Health, 70 （1）,158–164. https://doi.org/10.1080/07448481.2020.1728280

Murgia, E., Pera, M. S., Landoni, M., & Huibers, T. （2023, June）. Children on ChatGPT readability in an educational context: myth or opportunity?. In Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization （pp. 311-316）. https://doi.org/10.1145/3563359.3596996

OpenAI （2023）. OpenAI Official Website. Introducing ChatGPT - Learn more. Available online at: https://openai.com/blog/chatgpt/ （accessed February 10, 2023）.

Patel, N., Nagpal, P., Shah, T., Sharma, A., Malvi, S., & Lomas, D. （2023）. Improving mathematics assessment readability: Do large language models help?. Journal of Computer Assisted Learning, 39 （3）, 804-822. https://doi.org/10.1111/jcal.12776

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. （2011）. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532-1543.

Smith, D. R. （1989）. The Lexile Scale in Theory and Practice. Final Report.

Sakketou, F., & Ampazis, N. （2020）. A constrained optimization algorithm for learning GloVe embeddings with semantic lexicons. Knowledge-Based Systems, 195, 105628. https://doi.org/10.1016/j.knosys.2020.105628

Schwarm, S. E., & Ostendorf, M. （2005, June）. Reading level assessment using support vector machines and statistical language models. In Proceedings of the 43rd annual meeting of the Association for Computational Linguistics （ACL’05）, 523-530.

Schleicher, A., Zimmer, K., Evans, J., & Clements, N. （2009）. PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science. OECD Publishing （NJ1）.

Thorndike, E. L. （1921）. The teacher’s word book. New York: Teachers College, Columbia University

Weng, S. Y., Tseng, H. C., Sung, Y. T., & Chen, B.

（2019, October）. 基於階層式編碼架構之文本可讀性預測 （A Hierarchical Encoding Framework for Text Readability Prediction）. In Proceedings of the 31st Conference on Computational Linguistics and Speech Processing （ROCLING 2019） （pp. 334-342）.

Vajjala, S., & Meurers, D.（2012, June）. On improving the accuracy of readability classification using insights from second language acquisition. In Proceedings of the seventh workshop on building educational applications using NLP （pp. 163-173）.https://doi.org/10.3115/v1/D14-1162

Wang, Q., Wang, P., & Chang, Y. （2023）. Deep Learning-Based Intelligent Image Recognition and Its Applications in Financial Technology Services. Traitement du Signal, 40 （2）. https://doi.org/10.18280/ts.400233

Wang, S., Zeng, Q., Ni, W., Cheng, C., & Wang, Y. （2023）. ODP-Transformer: Interpretation of pest classification results using image caption generation techniques. Computers and Electronics in Agriculture, 209, 107863. https://doi.org/10.1016/j.compag.2023.107863

（N.d.）. Project Gutenberg. https://www.gutenberg.org/

A Comparative Study of Multi-document Summarization Techniques

Anushiya Thevapalan
University of Moratuwa, Sri Lanka
anushiya.22@cse.mrt.ac.lk

Nisansa de Silva
University of Moratuwa, Sri Lanka
nisansadds@cse.mrt.ac.lk

Abstract

Multi-document summarization (MDS) is an approach to extracting a concise and coherent summary of information from multiple source documents. This study presents a comparative analysis of MDS techniques, showcasing the progress in the field. Various MDS techniques are analyzed, and their strengths and weaknesses are compared, providing readers with insights to guide their own research directions. Additionally, benchmark datasets and standard evaluation techniques are presented. The experimental results highlight the variability in model performance across different datasets. For instance, the transformer-based PRIMERA model does well on the Multi-News dataset with a ROUGE-1 score of 42.0 but performs less effectively on others. In contrast, the PEGASUS model is more consistent across datasets, while the LED model excels on the BigSurvey-MDS and MS² biomedical datasets. The graph-based model HETERDOCSUMGRAPH outperforms the transformer-based model on the Multi-News dataset with a ROUGE-1 score of 46.05. The HGSum model, which combines transformer and graph techniques, performs best on the Multi-News dataset with a ROUGE-1 score of 50.64. These findings provide a clear overview of the current MDS techniques, highlighting their strengths and effectiveness in different areas.

1 Introduction

Document summarization is a fundamental task in natural language processing that aims to capture the essential information from the large text while preserving the overall meaning. Although single-document summarization has reached a level of maturity, Multi-Document Summarization (MDS) continues to pose significant challenges in the field of Natural Language Processing (NLP). The complexity arises from the need to amalgamate information from various sources, often characterized by conflicting, duplicate, or complementary details (Ma et al., 2022; Ferreira et al., 2014).

Document summarization techniques can be broadly categorized into three types: abstractive summarization, extractive summarization, and hybrid summarization (Afsharizadeh et al., 2022).

1.1 Abstractive Summarization

Abstractive summarization entails generating fresh phrases and sentences that are not directly copied from the source documents yet effectively convey their meaning. This approach to summarization necessitates a thorough comprehension of the text and the capability to produce original sentences that accurately encapsulate the key concepts and information present in the source documents (Figure 1) (Allahyari et al., 2017).

1.2 Extractive Summarization

On the contrary, extractive summarization focuses on choosing and merging significant sentences or phrases from the source documents to create the summary (Allahyari et al., 2017). This method involves identifying the crucial information within the source documents and selecting the most relevant sentences or phrases to construct the final summary (Figure 1). The specific techniques employed for extractive summarization may vary based on the length and complexity of the source documents. For instance, lengthier sources may require advanced techniques to identify the most vital sentences, whereas shorter sources can be summarized more straightforwardly by extracting keywords or phrases (Allahyari et al., 2017).

1.3 Hybrid Summarization

Hybrid summarization is an amalgamation of abstractive and extractive summarization methods (Ma and Zong, 2022; Afsharizadeh et al., 2022). As shown in Figure 1, this approach utilizes abstractive and extractive techniques to produce the final

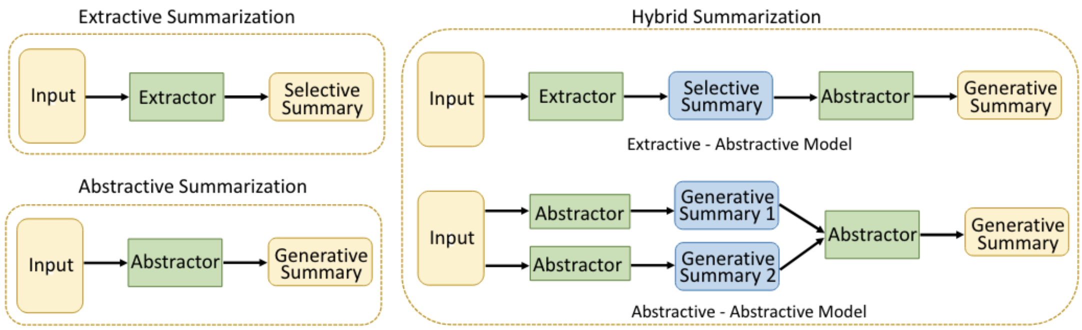


Figure 1: Summarization Construction Types for Text Summarization (Ma et al., 2022)

summary. To illustrate, the system may employ extractive techniques to identify the most crucial sentences or phrases from the source documents and subsequently utilize abstractive techniques to alter or rephrase these sentences for the final summary formation.

Given the diverse approaches to summarization, it is essential to evaluate and compare their effectiveness, particularly in the context of MDS, where challenges can be amplified. We are conducting a comparative analysis of multi-document summarization models to evaluate the performance of transformer-based and graph-based approaches. While transformer-based models have gained significant attention and shown strong results in various summarization tasks, there has been limited exploration of graph-based models in this context. Notably, no previous research has performed a comparative analysis that includes graph-based models alongside their transformer-based counterparts. By including these graph-based techniques in our evaluation, we aim to fill this gap in the literature, providing insights into how different modeling approaches perform across diverse datasets and applications. This analysis will not only enhance our understanding of the strengths and weaknesses of each type of model but also contribute to the development of more effective summarization solutions.

2 Related work

This section presents an overview of the models and techniques employed in this study. It highlights the work carried out by others in various areas relevant to this study.

2.1 Transformer based Methods

PRIMERA (Xiao et al., 2021) is a pre-trained model designed explicitly for Multi-Document Summarization (MDS), extending the capabilities of the Longformer Encoder-Decoder (LED) architecture. The model is pre-trained using a dataset called NewSHead (Gu et al.), which consists of 369,940 clusters of news articles covering similar topics. One of the key innovations in PRIMERA is the use of an Entity Pyramid strategy for generating synthetic summaries during pre-training. This approach identifies essential sentences by analyzing the frequency and relevance of named entities across documents. Through this method, PRIMERA can enhance the quality of summaries by prioritizing the most informative content from multiple sources. Experimental results demonstrate that PRIMERA significantly outperforms earlier MDS models across various benchmarks, showcasing its effectiveness in handling diverse and large-scale multi-document inputs (Xiao et al., 2021).

PEGASUS (Zhang et al., 2020) is a sequence-to-sequence model with gap-sentence generation as a pretraining objective tailored for abstractive text summarization. The key innovation in PEGASUS is its unique pre-training objective, where the model learns to predict entire sentences that have been masked from a document rather than individual tokens, as is common in traditional pre-training approaches like BERT (Zhang et al., 2020). This "gap-sentence generation" (GSG) technique allows PEGASUS to understand better a document's overall structure and key content, which is critical for generating coherent and informative summaries. PEGASUS has been shown to outperform other state-of-the-art models on multiple abstractive summarization benchmarks, particularly excelling in tasks that require summarizing long-form texts (Zhang et al., 2020)

DAMEN (Moro et al., 2022) is a model specifically designed to handle the challenges of multi-document summarization in the medical field. Unlike standard transformer models that may truncate or overlook critical information, DAMEN addresses this by using a discriminative approach to highlight essential content from clusters of related documents. It leverages token probability marginalization to ensure the generated summaries capture the most relevant details. This method proves particularly effective in the medical domain, where each piece of information is potentially vital (Moro et al., 2022). This approach showcases the potential for tailored neural models in domains requiring high precision, such as healthcare, ensuring the produced summaries are both accurate and informative.

SKT5SciSumm (To et al., 2024) is an Extractive-Generative approach for multi-document scientific summarization. The approach addresses the challenge of processing long and complex scientific texts by combining extractive and abstractive summarization techniques. SKT5SciSumm first utilizes the SPECTER model, which provides sentence-level embeddings trained on scientific texts, to generate dense representations of the input sentences. They then apply K-means clustering to extract the most important sentences from a collection of documents (To et al., 2024). These extracted sentences are passed to the T5 model, a state-of-the-art generative language model, which generates abstractive summaries. This combination balances extracting key information with generating fluent and coherent summaries (To et al., 2024).

2.2 Graph based Methods

In recent years, graph-based approaches have become popular in extractive summarization research. These techniques represent sentences, phrases, or words as nodes in a graph, with the connections between them serving as edges. By analyzing and scoring this network, the models identify and select the most informative and representative sentences to form the summary (Wang et al., 2020; Chen et al., 2021).

Graph-based methods construct graphs of sentences that are part of the document collection. The sentences make the graph's nodes, and edges are either drawn based on the similarity between sentences fulfilling the threshold criteria or belongingness to the same document (Wang et al., 2020; Jiang et al., 2022; Lu et al., 2022). Voting of neighboring nodes selects sentences to generate a summary. In the initial stages of graph-based methods for EDS tasks, researchers primarily utilized similarity scores between sentences in unsupervised ways, employing techniques such as TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2004b). The LexPageRank algorithm is based on eigenvector centrality to determine significant sentences, as was done successfully in the Google PageRank algorithm (Erkan and Radev, 2004a).

Data mining techniques have been employed to explore multi-document text summarization. (Baralis et al., 2013) utilized Association Rule Mining within the context of data mining to assess the outcomes of their summarization process. They introduced the GRAPHSUM algorithm to identify correlations among multiple terms in graph-based summarization. The Apriori algorithm was employed for association rule mining to identify correlations among terms, followed by the use of PageRank (Brin and Page, 2012) to rank salient sentences.

Canhasi (2017) introduced a method centered on a Five-Layered Heterogeneous Graph and Universal Paraphrastic Embeddings for query-focused extractive multi-document summarization. This research emphasizes relationships at both the sentence and document levels, incorporating aspects such as part-of-sentence similarity and query-to-sentence similarity (Canhasi, 2017).

Hierarchical graph structures also gained interest in recent research. HAHSum (Jia et al., 2020) is a hierarchical graph designed to address semantic sparsity by leveraging named entities. The HAHSum model incorporates three types of nodes: named entity nodes, word nodes, and sentence nodes. The named entity nodes are represented as anonymized tokens. The graph construction process is as follows: word nodes are connected to the corresponding sentence node with directed edges if they appear within the same sentence. Two named entity nodes are connected by undirected edges if they refer to the same entity, while two sentence nodes are connected by undirected edges if they share a trigram. Furthermore, sequentially occurring words and entities are connected with directed edges. This approach effectively captures implicit information in an explicit manner, allowing for a more comprehensive encoding of the input data (Jia et al., 2020).

HETERDOCSUMGRAPH (Wang et al., 2020) represents cutting-edge developments in graph-based neural networks for summarization tasks. It utilizes heterogeneous graph neural networks to enhance extractive summarization. This model constructs a graph where nodes represent sentences and edges capture various relationships such as semantic similarity or document structure (Wang et al., 2020). By leveraging these graph structures, HETERDOCSUMGRAPH can identify and extract the most relevant sentences from multiple documents, ensuring that the resulting summary represents the source material. This graph-based approach allows for a more nuanced understanding of the interconnections between different pieces of information, which is crucial for effective multi-document summarization (Wang et al., 2020).

The development of graph-based summarization techniques marks significant progress in how text is analyzed and summarized. By using graph structures to represent relationships between sentences, words, and entities, these methods provide a strong foundation for creating summaries that are both informative and accurate. As research explores new ways of building and analyzing these graphs, these techniques are expected to become even more critical in creating advanced summarization models that can effectively manage the complexity of today's data.

2.3 Hybrid Methods

CovSumm (Karotia and Susan, 2023) is an unsupervised hybrid summarization model developed for the CORD-19 dataset, which contains COVID-19-related scholarly articles. By combining transformer-based and graph-based methodologies, the model effectively generates summaries from a large volume of scientific texts. The transformer leverages deep learning for contextual understanding, while the graph-based approach enhances the representation of relationships within the documents. This dual strategy enables CovSumm to produce coherent, concise, and informative summaries, making it a valuable resource for researchers and healthcare professionals seeking critical insights from extensive literature.

HGSUM (Li et al., 2023) — an MDS model that extends an encoder-decoder architecture to incorporate a heterogeneous graph to represent different semantic units (e.g., words and sentences) of the documents. To preserve only key information and relationships of the documents in the heterogeneous graph, HGSUM uses graph pooling to compress the input graph (Li et al., 2023).

Khaliq et al. (2024) introduces a framework that integrates topic-aware heterogeneous graph neural networks (HGNN) with transformer models to improve the abstractive summarization of medical scientific documents. This model utilizes HGNN to capture complex relationships among various topics while employing transformer architecture for effective sequence modeling.

3 Datasets

The process of choosing datasets for this study included evaluating their popularity and novelty. We selected datasets from various domains, taking into account the type of data, whether it was short, long, or a hybrid of both document types. Table 4 in the appendix shows the summary of datasets used for evaluation.

The Multi-News dataset (Fabbri et al., 2019) constitutes a comprehensive compilation of news articles and their corresponding human-generated summaries sourced from the website newser.com (Abid, 2022). These summaries are crafted by professional editors and incorporate links to the original articles. One of the most defining characteristics of the Multi-News dataset is its breadth. It incorporates articles from over 1,500 distinct news sites, making it one of the most diverse datasets in the domain of news summarization. It covers various topics, including politics, economics, science, entertainment, etc.

The Multi-News dataset better captures different news organizations' varied styles, tones, and editorial slants, providing a more challenging and realistic task for summarization models. Approximately 85% of the summaries were written by a group of 20 editors, ensuring that the dataset maintains a high level of coherence and consistency across summaries. This contrasts with datasets such as WikiSum (Liu et al., 2018), where the summaries are crowd-sourced or auto-generated from scraped web content, resulting in less uniformity in quality and style. The involvement of professional editors also introduces a diversity of editorial perspectives, as each editor may prioritize different elements of the articles they summarize. This ensures that the dataset encompasses a variety of summarization approaches, which can lead to more robust evaluations of multi-document summarization models. (Abid, 2022).

The Multi-XScience dataset (Lu et al., 2020) is a distinctive compilation formed by merging papers from arXiv.org and Microsoft Academic Graph (MAG) (Sinha et al., 2015) to establish pairs of target summaries and multi-reference documents. This dataset plays a vital role in addressing the complexities of summarizing scientific literature by pairing target summaries with multiple reference documents from these repositories, allowing for generating summaries that synthesize information across various related papers (Lu et al., 2020).

A vital characteristic of the Multi-XScience dataset is its scale. The dataset consists of 1.3 million arXiv papers. The dataset underwent a curation process involving the cleaning of 1.3 million arXiv LATEX files and aligning them with MAG references. This process was refined through five iterations of cleaning, followed by human validation to ensure data quality (Lu et al., 2020).

The dataset covers various scientific disciplines, including physics, computer science, and biology, making it highly interdisciplinary. This broad scope allows researchers to evaluate how well summarization models generalize across different fields (Lu et al., 2020).

Liu et al. (2018) introduced the WikiSum dataset; generated from summarizing the long Wikipedia articles. It provides a large-scale, naturally occurring dataset. It is built by scrapping the web and designed to address the challenges in long input sequences in multi-document summarization. The dataset includes 1.5 million documents paired with their summaries (Liu et al., 2018).

The BigSurvey dataset (Liu et al., 2023) is introduced in the context of generating structured summaries from large collections of academic papers. This dataset is designed to address the specific challenges associated with summarizing extensive scientific literature. These summaries capture key contributions, methodologies, and results across multiple papers, offering a more organized and systematic representation of the underlying academic content (Liu et al., 2023).

A notable feature of BigSurvey is its emphasis on hierarchical structure, making it well-suited for tasks where organized academic discourse is essential (Li et al., 2023). The summaries are designed to reflect the structure of academic reviews, clearly distinguishing between sections like methods and results—something not addressed in datasets like Multi-News (Fabbri et al., 2019; Li et al., 2023).

Rotten Tomatoes, a website categorizing film reviews, has been widely used in sentiment analysis research. The Rotten Tomatoes dataset (Leone, 2020) includes movie reviews and meta-reviews, with meta-reviews created by professional editors and accompanied by a Tomatometer score, which reflects overall critic reception. This dataset is a key resource for studying sentiment analysis and summarization in the movie review domain.

The Wikipedia Current Events Portal (WCEP) Multi-Document Summarization dataset is a large-scale resource designed to support research in multi-document summarization (MDS) by leveraging a diverse range of real-world events from Wikipedia (Ghalandari et al., 2020). This dataset consists of news articles curated from the Wikipedia Current Events Portal, focusing on global news stories from various domains, such as politics, science, and social events. Each cluster of documents covers a specific event or topic, providing multiple perspectives and details that need to be summarized into a coherent and concise summary (Ghalandari et al., 2020).

4 State-of-the-art Models

We conducted an extensive literature review to identify leading models in multi-document summarization (MDS). From the range of advanced models available, we selected PRIMERA (Xiao et al., 2021), PEGASUS (Zhang et al., 2020), and LED (Beltagy et al., 2020) to represent transformer-based approaches, alongside MGSum (Jin et al., 2020), GraphSum (Li et al., 2020), and HETERDOCSUMGRAPH (Wang et al., 2020) for graph-based models. Additionally, we included HGSum (Li et al., 2023), a hybrid model that combines the strengths of both transformer and graph-based architectures. The selection criteria were based on performance metrics, the year and venue of publication, and the distinct strategies these models employ in processing multi-document inputs. Together, these models encapsulate the current advancements in summarization techniques and provide a broad overview of contemporary methodologies in MDS. Table 3 in the appendix summarises the number of parameters and length of the input and output for the chosen models.

PRIMERA (Pyramid-based Masked Sentence Pre-training for Multi-document Summarization) by Xiao et al. (2021) is noted for its robust performance in handling large-scale document inputs. It leverages a pyramid-based approach to generate summaries by identifying and utilizing hierarchical structures within the text. PRIMERA's ability to manage complex information from multiple sources makes it a strong candidate for this comparative analysis.

PEGASUS (Zhang et al., 2020) employs a novel pre-training objective called gap-sentence generation, which has shown state-of-the-art results in various summarization datasets. This model is particularly noted for its strong abstractive summarization capabilities, making it a critical inclusion for evaluating high-quality summarization (Zhang et al., 2020). Similarly, the Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) model extends the BERT and RoBERTa architectures to handle longer sequences, addressing the limitations of standard Transformer models in processing long documents effectively.

The Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) model, addresses the challenge of handling long sequences by extending the Transformer architecture. LED uses a combination of local and global attention mechanisms to process long documents without the quadratic increase in computational complexity typical of traditional Transformers (Beltagy et al., 2020). This model is particularly effective in multi-document summarization as it can seamlessly integrate information across extensive texts, ensuring that the summary captures the breadth of content from the input documents (Beltagy et al., 2020).

HETERDOCSUMGRAPH (Wang et al., 2020) utilizes heterogeneous graph neural networks to enhance extractive summarization. This model constructs a graph where nodes represent sentences and edges capture various relationships such as semantic similarity or document structure (Wang et al., 2020). By leveraging these graph structures, HETERDOCSUMGRAPH can identify and extract the most relevant sentences from multiple documents, ensuring that the resulting summary represents the source material. This graph-based approach allows for a more nuanced understanding of the interconnections between different pieces of information, which is crucial for effective multi-document summarization (Wang et al., 2020).

HGSum (Compressed Heterogeneous Graph for Abstractive Multi-Document Summarization) (Li et al., 2023) builds on the graph-based approach by incorporating a hierarchical structure. HGSum introduces a transformer-based approach combining the benefits of graphs. It represents documents at multiple levels of granularity, allowing the model to integrate and synthesize information from various layers (Li et al., 2023). This hierarchical approach ensures that both fine-grained details and high-level summaries are considered, resulting in a comprehensive and coherent summary. HGSum's ability to process and summarize large and complex document sets makes it a strong contender in the field of multi-document summarization (Li et al., 2023).

Including these models in a comparative analysis allows for a thorough examination of the strengths and weaknesses of different approaches to multi-document summarization. This analysis provides a comprehensive overview of the current state-of-the-art in the field by evaluating models that use sparse attention mechanisms, novel pre-training objectives, extended Transformer architectures, and graph-based neural networks. Each model's unique approach to handling the challenges of summarizing multiple documents contributes valuable insights into integrating best and distilling information from diverse sources.

5 Evaluation metrics

ROUGE, which stands for Recall-Oriented Understudy for Gisting Evaluation, comprises both a set of metrics and a software package employed in the assessment of automatic summarization and machine translation software within natural language processing. These metrics compare automatically generated summaries or translations and a reference (or set of references) provided by humans (Lin, 2004).

BLEU- Bilingual Evaluation Understudy: It employs precision-based metrics to evaluate the similarity between machine-generated and reference translations. The evaluation considers the machine-generated text's fluency (grammatical correctness) and adequacy (semantic equivalence). BLEU is widely utilized in the field of natural language processing and machine translation to quantify the effectiveness of translation models and algorithms. It is used in (Christensen et al., 2013; Tzouridis et al., 2014; ShafieiBavani et al., 2016).

Other evaluation metrics used are Precision, Recall, F-measure, Average Continuity, Pyramid, Correlation coefficients, Amazon mTurk (AMT), etc.

We selected ROUGE for evaluation due to its widespread use in natural language processing, making it easier to compare our results with previous studies. Its broad acceptance ensures consistent benchmarking across various summarization models. By aligning our findings with earlier research, ROUGE enhances the reliability and validity of our comparative analysis.

Datasets	Metric	PRIMERA [1]	PEGASUS [2]	LED [3]	HETERDOCSUMGRAPH [4]
Multi-News [5]	ROUGE-1	42.0 [1]	32.0 [1]	17.3 [1]	46.05 [4]
	ROUGE-2	13.6 [1]	10.1 [1]	3.7 [1]	16.35 [4]
	ROUGE-L	20.8 [1]	16.7 [1]	10.4 [1]	42.08 [4]
Multi-XScience [6]	ROUGE-1	29.1 [1]	27.6 [1]	14.6 [1]	32.56
	ROUGE-2	4.6 [1]	4.6 [1]	1.9 [1]	7.29
	ROUGE-L	15.7 [1]	15.3 [1]	9.9 [1]	17.90
WikiSum [7]	ROUGE-1	28.0 [1]	24.6 [1]	10.5 [1]	30.48
	ROUGE-2	8.0 [1]	5.5 [1]	2.4 [1]	10.04
	ROUGE-L	18.0 [1]	15.0 [1]	8.6 [1]	21.79
BigSurvey-MDS [8]	ROUGE-1	23.9 [11]	38.9 [8]	39.8 [8]	37.26
	ROUGE-2	4.1 [11]	9.0 [8]	9.4 [8]	8.05
	ROUGE-L	11.7 [11]	16.2 [8]	16.1 [8]	15.91
MS^2 [9]	ROUGE-1	12.8 [11]	12.7 [11]	25.8 [11]	23.76
	ROUGE-2	2.0 [11]	1.5 [11]	8.4 [11]	6.92
	ROUGE-L	8.1 [11]	8.3 [11]	19.3 [11]	17.85
Rotten Tomatoes [10]	ROUGE-1	25.4 [12]	27.4 [12]	25.6 [12]	-
	ROUGE-2	8.4 [12]	9.5 [12]	8.0 [12]	-
	ROUGE-L	19.8 [12]	21.1 [12]	19.6 [12]	-
WCEP [13]	ROUGE-1	43.11 [14]	42.43 [14]	43.05 [14]	-
	ROUGE-2	21.85 [14]	17.33 [14]	20.94 [14]	-
	ROUGE-L	35.89 [14]	32.35 [14]	34.99 [14]	-

Table 1: ROUGE Scores of Different Models on Different Datasets. The sources are as follows.
[1] Xiao et al. (2021), [2] Zhang et al. (2020), [3] Beltagy et al. (2020), [4] Wang et al. (2020), [5] Fabbri et al. (2019), [6] Lu et al. (2020), [7] Liu et al. (2018), [8] Liu et al. (2023), [9] DeYoung et al. (2021), [10] Leone (2020), [11] Hewapathirana et al. (2023), [12] DeYoung et al. (2023) [13] Ghalandari et al. (2020) , [14] Li et al. (2023)

6 Experimental results and discussion

This section presents the findings of our study, which compares the performance of state-of-the-art models on different datasets from various domains. Specifically, we report on the performance of these models using ROUGE (Lin, 2004) scores. We also examine the impact of dataset characteristics, such as the number of documents and documents per cluster, on the performance of the models.

We gathered results from previous studies that utilized the same models and datasets. In cases of conflicting findings, we prioritized the most recent study and the conference where it was presented. We also ensured that the original parameters specified in the studies introducing the models were maintained.

In our research, we specifically evaluated the performance of the HETERDOCSUMGRAPH model (Wang et al., 2020) on Multi-XScience (Lu et al., 2020), MS^2 (DeYoung et al., 2021), and on the newly released BigSurveyMDS dataset (Liu et al., 2023). For this evaluation, we used the same parameters as those in the original HETERDOCSUMGRAPH setup and tested the model on the dataset's test set. The results were then compiled and summarized in Table 1 and 2 to facilitate comparison and analysis. To our knowledge, this is the first time the HETERDOCSUMGRAPH model has been evaluated on the Multi-XScience, MS^2, and BigSurvey MDS dataset.

The results from the two tables provide insights into the performance of various models on multiple datasets using the ROUGE evaluation metrics. Across different datasets, the models exhibit varying strengths, highlighting the diversity of approaches in handling multi-document summarization (MDS) tasks.

On the Multi-News dataset, HGSum demonstrated the highest ROUGE-1 score 50.64, significantly outperforming other models like HETERDOCSUMGRAPH 46.05, GraphSum 45.71,

Datasets	Metric	HGSum [1]	GraphSum [2]	MGSum [3]
Multi-News [4]	ROUGE-1	50.64	45.71	45.63
	ROUGE-2	21.69	17.12	16.71
	ROUGE-L	45.90	41.99	40.92
WCEP [5]	ROUGE-1	44.21	39.56	38.88
	ROUGE-2	21.81	14.38	14.22
	ROUGE-L	36.21	29.41	23.37

Table 2: ROUGE Scores of Different Models on Different Datasets. The sources are as follows. All the metrics are taken from [1]

[1] Wang et al. (2020), [2] Li et al. (2020), [3] Jin et al. (2020),
[4] Fabbri et al. (2019), [5] Ghalandari et al. (2020) , [6] Li et al. (2023)

and MGSum 45.63. The superior performance of HGSum can be attributed to its hybrid approach, which leverages both graph structures and transformers. Similarly, HGSum excelled in ROUGE-2 and ROUGE-L metrics, indicating its capacity to generate both accurate and coherent summaries from complex multi-document inputs. HETERDOCSUMGRAPH, which relies on graph-based methods, exhibited strong performance, particularly in handling hierarchical sentence relationships. However, transformer-only models such as PRIMERA and PEGASUS performed comparably lower on ROUGE metrics for this dataset.

On the WCEP dataset, HGSum again led the results with the highest ROUGE-1 score (44.21), followed by GraphSum (39.56). Interestingly, HETERDOCSUMGRAPH was not evaluated for this dataset, making it challenging to assess how well graph-based models perform across datasets that focus on different types of input documents. Despite this, the consistent success of hybrid models like HGSum suggests that combining graph structures and transformer models yields superior results when handling real-world, complex datasets.

PEGASUS and LED showed competitive performance when analyzing other datasets like Multi-XScience, WikiSum, and BigSurvey-MDS. For instance, PEGASUS achieved solid results on BigSurvey-MDS (ROUGE-1 of 38.9), though LED marginally outperformed it in ROUGE-1 and ROUGE-2 scores, likely due to its ability to process longer sequences more effectively. However, PRIMERA underperformed on BigSurvey-MDS compared to its performance on datasets like Multi-News and WikiSum, suggesting that some models may be domain-specific.

In the biomedical-focused MS² dataset, LED demonstrated superior performance (ROUGE-1 of 25.8), outperforming both PRIMERA and PEGASUS. This can be attributed to LED's local and global attention mechanism, which is particularly effective in summarizing longer, more detailed texts typical in medical literature. HETERDOCSUMGRAPH also showed strong results in other datasets, but its absence from MS² prevents direct comparison.

In conclusion, hybrid models like HGSum and graph-based models like HETERDOCSUMGRAPH generally outperform traditional transformer-based models across most datasets. This suggests that combining the strengths of both transformers and graph-based approaches can better capture complex relationships between multiple documents, resulting in more accurate and coherent summaries. However, the performance of models can vary significantly across datasets, indicating that dataset characteristics heavily influence model effectiveness.

7 Conclusion

Multi-document summarization (MDS) holds promise in transforming how we process large datasets, but it faces challenges like managing diverse documents, reducing redundancy, and maintaining coherence. This analysis underscores the importance of dataset characteristics in selecting summarization models, as performance varies significantly across domains. There is no universal solution for MDS, and future research should focus on creating more adaptive models capable of handling various dataset features, including document length and topic diversity. Additionally, hybrid approaches and advanced techniques such as transfer learning could improve model robustness and generalizability across different domains.

References

Azal Minshed Abid. 2022. Multi-document text summarization using deep belief network. *International Journal of Advances in Scientific Research and Engineering (IJASRE)*, 8(8):56–65.

Mahsa Afsharizadeh, Hossein Ebrahimpour-Komleh, Ayoub Bagheri, and Grzegorz Chrupała. 2022. A survey on multi-document summarization and domain-oriented approaches. *Journal of Information Systems and Telecommunication (JIST)*, 1(37):68.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Elena Baralis, Luca Cagliero, Naeem Mahoto, and Alessandro Fiori. 2013. Graphsum: Discovering correlations among multiple terms for graph-based summarization. *Information Sciences*, 249:96–109.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Sergey Brin and Lawrence Page. 2012. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833.

Ercan Canhasi. 2017. Query focused multi-document summarization based on five-layered graph and universal paraphrastic embeddings. In *Artificial Intelligence Trends in Intelligent Systems: Proceedings of the 6th Computer Science On-line Conference 2017 (CSOC2017), Vol 1 6*, pages 220–228. Springer.

Moye Chen, Wei Li, Jiachen Liu, Xinyan Xiao, Hua Wu, and Haifeng Wang. 2021. Sgsum: transforming multi-document summarization into sub-graph selection. *arXiv preprint arXiv:2110.12645*.

Janara Christensen, Stephen Soderland, Oren Etzioni, et al. 2013. Towards coherent multi-document summarization. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1173.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486*.

Jay DeYoung, Stephanie C Martinez, Iain J Marshall, and Byron C Wallace. 2023. Do multi-document summarization models synthesize? *arXiv preprint arXiv:2301.13844*.

Gunes Erkan and Dragomir Radev. 2004a. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 365–371.

Günes Erkan and Dragomir R Radev. 2004b. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Rafael Ferreira, Luciano de Souza Cabral, Frederico Freitas, Rafael Dueire Lins, Gabriel de França Silva, Steven J Simske, and Luciano Favaro. 2014. A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications*, 41(13):5780–5787.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. *arXiv preprint arXiv:2005.10070*.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, Hongkun Yu, You Wu, Cong Yu, Daniel Finnie, Jiaqi Zhai, and Nicholas Zukoski. Generating Representative Headlines for News Stories. In *Proc. of the the Web Conf. 2020*.

Kushan Hewapathirana, Nisansa De Silva, and C.D. Athuraliya. 2023. Multi-document summarization: A comparative evaluation. In *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*, pages 19–24.

Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. 2020. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3622–3631.

Ming Jiang, Yifan Zou, Jian Xu, and Min Zhang. 2022. Gatsum: graph-based topic-aware abstract text summarization. *Information Technology and Control*, 51(2):345–355.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6244–6254.

Akanksha Karotia and Seba Susan. 2023. Covsumm: an unsupervised transformer-cum-graph-based hybrid document summarization model for cord-19. *The Journal of Supercomputing*, 79(14):16328–16350.

Ayesha Khaliq, Atif Khan, Salman Afsar Awan, Salman Jan, Muhammad Umair, and Megat Farez Azril Zuhairi. 2024. Integrating topic-aware heterogeneous graph neural network with transformer model for medical scientific document abstractive summarization. *IEEE Access*.

Stefano Leone. 2020. Rotten tomatoes movies and critic reviews dataset.

Miao Li, Jianzhong Qi, and Jey Han Lau. 2023. Compressed heterogeneous graph for abstractive multi-document summarization. *arXiv preprint arXiv:2303.06565*.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging graph to improve abstractive multi-document summarization. *arXiv preprint arXiv:2005.10043*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2023. Generating a structured summary of numerous academic papers: Dataset and method. *arXiv preprint arXiv:2302.04580*.

Menghua Lu, Lijia Liang, and Gongshen Liu. 2022. Parallel relationship graph to improve multi-document summarization. In *International Conference on Artificial Neural Networks*, pages 630–642. Springer.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. *arXiv preprint arXiv:2010.14235*.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, 55(5):1–37.

Ye Ma and Lu Zong. 2022. Parallel hierarchical transformer with attention alignment for abstractive multi-document summarization. *arXiv preprint arXiv:2208.07845*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Davide Freddi. 2022. Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 180–189.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2016. On improving informativity and grammaticality for multi-sentence compression. *arXiv preprint arXiv:1605.02150*.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

Huy Quoc To, Hung-Nghiep Tran, Andr'e Greiner-Petter, Felix Beierle, and Akiko Aizawa. 2024. Skt5scisumm-a hybrid generative approach for multi-document scientific summarization. *arXiv preprint arXiv:2402.17311*.

Emmanouil Tzouridis, Jamal A Nasir, and Ulf Brefeld. 2014. Learning to summarise related sentences. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1636–1647.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. *arXiv preprint arXiv:2004.12393*.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. Primera: Pyramid-based masked sentence pre-training for multi-document summarization. *arXiv preprint arXiv:2110.08499*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

A Model parameter sizes

Model	#parameters	Len-in	Len-out
PRIMERA	447M	4,096	512
PEGASUS	568M	1,024	512
LED	459M	16,384	512
GraphSum	463M	4,050	300
HGSUM	501M	4,096	512
GPT-3	175B	2,049	512

Table 3: Model parameter sizes. Len-in and Len-out denote the maximum lengths of the model input and output, respectively (Li et al., 2023; Floridi and Chiriatti, 2020)

B Summary of datasets used for evaluation

Dataset	Total number of documents	Average number of documents per cluster	Domain
Multi-News (Fabbri et al., 2019)	56K [1]	3.5 [1]	News articles
Multi-Xscience (Lu et al., 2020)	40K [1]	2.8 [1]	Related-work section in scientific articles
Wikisum (Liu et al., 2018)	1.5M [1]	40 [1]	Wikipedia articles
BigSurvey-MDS (Liu et al., 2023)	430K	61.4	Human-written survey papers on various domains
MS^2 (DeYoung et al., 2021)	470K	23.5	Reviews of scientific publications in the medical domain
Rotten Tomato Dataset (Leone, 2020)	244K	26.8	Movie reviews
WCEP (Ghalandari et al., 2020)	650K	63	Human written summaries on news events

[1] (Xiao et al., 2021)

Table 4: Summary of datasets used for evaluation

Donation Intention Classifications Task in Non-Profit Organizations Based on Document Classification Techniques

Kai-Chi Yu
Department of Data Science,
Soochow University, Taiwan
10173209@gm.scu.edu.tw

Chin-Jung Wu
Department of Data Science,
Soochow University, Taiwan
wanzijks@gmail.com

Jheng-Long Wu
Department of Data Science,
Soochow University, Taiwan
jlwu@gm.scu.edu.tw

Abstract

In contemporary society, non-profits make up for what companies and the government lack. Fundraising and donor communication skills are crucial for success. This research uses natural language processing techniques to analyze donation messages, employing BERT embeddings combined with SVM, XGBoost, LLM model, and Voting Classifier. Results show that BERT significantly improves accuracy. The study recommends expanding the dataset and incorporating different platforms to enhance model precision, optimize fundraising strategies, and improve donor interactions, ultimately encouraging social contributions.

Keywords: Non-Profit Organizations, Natural Language Processing, BERT, Intent Detection

1 Introduction

In recent years, non-profit organizations (NPOs) have carried out significant social missions; their operating funds mainly come from donations from the public, forming "public funds" to support various initiatives and humanitarian aid work.

In an era of rapid development of artificial intelligence, investigating how to employ technology to increase fundraising efficiency has become a new challenge for NPOs. Tseng (2023) using precise text analysis on NPOs, can create targeted and successful fundraising strategies by better understanding the needs and intents of donors. Therefore, this research selects donation messages in online donations on NPO official websites as the research topic. This study develops a tool to automatically classify donation messages so that NPOs can automatically analyze many donation messages, accurately capture donors' concerns, needs, and opinions, and provide effective fundraising strategies and communication methods. To begin with, data pre-processing on donation messages will be performed to remove noise and perform multi-label classification according to classification standards. The pre-trained BERT, Support Vector Machine (SVM), XGBoost, and LLM model in natural language processing (NLP) for feature extraction and classification. After training, the model's prediction performance was evaluated based on the Macro-average F1 score. Secondly, it is hoped that this research can provide NPOs with tools to process and analyze many donation messages efficiently. In addition, the research also established a rich data set of donor message intentions, laying the foundation for future model training and optimization. We expect that these tools and methods can help NPOs more effectively understand and respond to the needs of donors, provide more precise goals and directions for future fundraising strategies, leading to more effective social services and welfare provision.

2 Related Work

2.1 Research on Donations to NPO

In recent years, the COVID-19 pandemic has greatly impacted the charity sector in various countries, and Taiwan has not been immune. In the post-epidemic era, it is crucial for NPOs to increase the frequency of continuous donations. Research shows that for every 1% increase in fundraising, donations will decrease by 0.84%. Therefore, reducing the churn rate and increasing the "renewal rate" play an important role in the long-term stable development of the organization (Marudas et al., 2012). Donation behavior is affected by many factors, including age, income, gender, region, and education level. Research shows that economic status is a key factor, with donation amounts positively correlated to annual income (Wang, 2023). Not only that, according to statistics, men's total donation amount is relatively high, but women's donation willingness and frequency are higher. Therefore (Chen, 2014) believes that gender may be a potential factor in developing donation strategies.

In addition to introductions from relatives and friends, advertising, and tax saving, Yang (2023) pointed out empathy can trigger altruistic donation behavior. Hsiao (2020) discovered in his research that there is a significant positive correlation between the convenience of contribution and willingness to give. Highlights the value of diversified donation methods and channels.

As Taiwan enters an aging society, middle-aged and elderly people have become the main donor group. Nonetheless, NPOs will eventually need to cultivate potential donors from Generation Z. Targeting Generation Z donors, gifts and rewards are used to encourage donations and marketing techniques for placing fundraising gifts. These flash marketing strategies help attract cooperation and strengthen corporate social responsibility (Hsueh, 2014).

2.2 Message Analysis in Social Media Applications

With the rapid development of Internet technology, social media has become an indispensable platform for information exchange. NPOs can use these platforms to strengthen information dissemination, interaction, fundraising, and volunteer management, increasing public awareness, support, resource collection, and social influence. Facebook messages have become important for governments, businesses, and NPOs to understand users' emotions and opinions. According to Kuang (2015) research, message analysis can identify opinion leaders with a 70% chance of success and can discern between positive, neutral, and negative emotions with a 90% accuracy rate; Lee (2023) noted that LINE's ecology makes information transmission more convenient. Nonetheless, the age group of LINE users is relatively high, prompting NPOs to focus on engaging younger potential donors. In contrast, Instagram attracts many young users and combines pictures and short videos to become an important social media platform. Lin (2020) found that users often use emojis to express their feelings and appreciation of food.

Overall, NPOs need to understand the characteristics and limitations of each platform and formulate appropriate fundraising strategies to promote fundraising activities and improve donation efficiency.

2.3 Application of Deep Learning and Machine Learning in Sentiment Analysis

With the rapid development of deep learning and machine learning technologies, text classification and sentiment analysis fields have advanced dramatically in recent years. BERT (published by Google in 2018) has brought breakthrough development in sentiment analysis. Based on the Transformer architecture, it can consider the contextual information of words at the same time and achieve good results when applied to multi-category, multi-label, and multi-output classification. In addition, machine learning has demonstrated excellent capabilities in sentiment analysis problems. Support vector machine (SVM), as a powerful supervised learning model, is often used to solve classification and regression analysis problems. Lin (2020) used SVM to predict the rise and fall of the foreign exchange market using news headlines, showing its successful application in financial markets. Another machine learning algorithm, Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2014), is an efficient gradient-boosting tree algorithm. Huang & Wang (2023)'s research used the post data set of China's Sina Weibo during the COVID-19 period for sentiment analysis, utilizing XGBoost to determine the correlation between the word vector sparse matrix and the user's mental health status. The findings indicate that the epidemic has an extremely negative impact on people's psychology.

This study further explores the application and advantages of voting classifiers. Integrating multiple models improves performance while addressing bias and overfitting issues. Yang (2022) applied a voting classifier in Taiwan stock market futures trading to predict the closing price trend of Taiwan stock futures, which improved transaction efficiency. Based on the above, machine learning and deep learning have boundless promise in the NLP domain. They can offer more precise perceptions of people's feelings and actions, serving as valuable references. This research aims to develop an automatic message classifier to classify donation messages into ten categories. The tool will enable fast, accurate analysis of donation messages, supporting fundraising strategies across different contexts while capturing and analyzing social and emotional trends and behavioral intentions.

2.4 Large Language Models for Text Classification

Although large language models (LLM) have made significant progress in natural language processing, especially with zero-shot and few-shot tasks (Sun et al., 2023), LLM was used to predict donation intention by classifying challenged benchmarks. This is consistent with previous difficulties encountered using LLM in structured or hierarchical tasks, where the complexity and nuances of the label space constitute limitations. Autoregressive-based LLMs like GPT (Touvron et al., 2023; OpenAI, 2024) generate text one token at a time conditioned on previous inputs. Despite LLMs being powerful API-accessible tools that let you make heuristic-based hints without any training task-specific, they may not always get accurate classification, specifically in more complex or less sophisticated tasks. Sun et al. (2023), LLM does well in flat classification problems. Still, it is unsuitable for tasks with highly structured label spaces, such as the hierarchical donation intention classification task used in this study. Said results suggest that, while LLM is a promising approach for some NLP tasks, the structure of labels in more complex or nuanced contexts could demand further tuning the need to complement the classification model with other methods.

3 Method

This study proposes to create a dataset for the donation intent classification, which is collected from donation websites for NPOs, and define the label of donation intent by human annotation. Finally, we apply several machine learning and deep learning models to train and predict the donation intent. The research framework is shown in Figure 1.

3.1 Dataset

In this section, we describe the process of collecting donation information from the donation database of NPOs official website. These messages reflect donors' remarks, feedback, and needs during the donation process, typically showing support for organizations, events, or service projects. The messages range in content and length and contain emojis, Chinese, and English. Some may be brief expressions of gratitude, while others can be more in-depth and sentimental assessments. The data collection covers 2020/01 to 2023/07. The original data from the NPOs official website contains 22,994 messages.

3.2 Data Preprocessing

In this study, following the collection of donation messages from NPOs, we manually processed the data to ensure the quality of the training set and consistency in feature extraction. This involved reviewing each message and removing irrelevant characters, duplicate content, and unusual conditions to ensure data consistency and accuracy; we carried out the following data noise-cleaning tasks: (1) Emoji emoticons and special symbols: They will be deleted if there is no semantic change or is redundancy. (2) HTML special tags and syntax, such as <br/>, to prevent sentence fragmentation or different semantics. (3) Punctuation: Delete as it adds nothing to the message's original meaning. (4) Duplicate messages: Removed to avoid negatively impacting the training process. While emoticons often convey emotional intent, we removed these symbols to maintain consistency and clarity in our analysis, especially when they were used to repeat the emotional expression of an existing message. After data processing, 4717 messages were finally

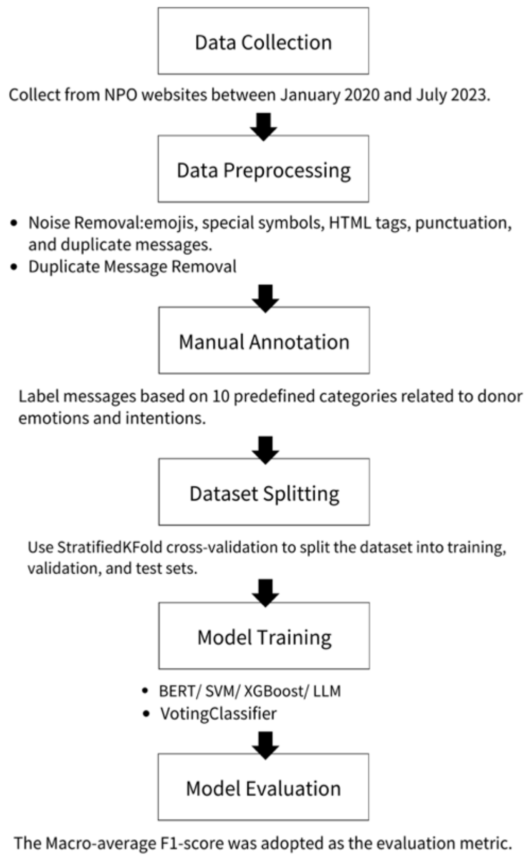


Figure 1. Research Process Flowchart

Categories Type	Definition	Example
Support Again	Supporters who have donated may clearly indicate their continued support for the organization in their messages, or update the authorization information due to changes in card information.	"I have donated before and am willing to continue to support."
Environmental Protection	Donors' actual actions and expectations for environmental protection, may want to switch from paper to electronic forms to expressing their support for environmental protection through digital channels.	"No need for paper DM"
Personal Information	Donors are concerned about privacy protection and ensure that donors' personal and financial flow information are properly protected to prevent data leakage or improper use, and whether the donation will be made public. (Those who leave their name, address or phone number in the donation message also fall into this category.)	"Who am I? What is my phone number and company name?"
Taxation	Donors want to know whether their donations are eligible for tax deductions and how to obtain relevant tax receipts or documents. There are also issues related to tax savings.	"Can this donation be used for tax filing?"
Giveaways	When people donate to charities, some organizations will provide gifts in full to express their gratitude and encourage for the donation. Donors may have questions about the gift.	"I want to ask for a thank you gift" or "I don't need any gifts"
Gratitude	Donors may express their gratitude and appreciation to the charity and praise it for supporting rural and disadvantaged groups or share their own experiences of benefiting from the organization.	"I sincerely thank you for your work, you did a great job!"
Encouragement and expectations	Donors may leave messages to express their support and expectations for the charity and full of expectations for the organization's future and encourage the organization to make greater contributions to social welfare.	"Thank you for your hard work! Come on!"
Care and blessings	Donors may express blessings and good wishes in their messages, hoping that underprivileged groups would be able to overcome obstacles and lead fulfilling lives.	"I pray for these children that their lives will be filled with joy and success."
Hope	Through words of encouragement, donors convey positive energy and encourage underprivileged groups to face challenges, establish and achieve success.	"Seeing the hard work of children in rural areas, I hope my donation can bring them more opportunities."
Empathy	Donors convey empathy, identification, and understanding with underprivileged populations in their communications. They will also express their recognition for the charity's development plans.	"I have also faced difficulties, and I understand their situation."

Table 1. The Definition and Examples of Ten Donation Intentions.

selected for this study's data set. Using the labeling criteria to complete the manual token samples, the statistics show that 4192 messages belong to 1 category, 502 messages belong to 2 categories, and 23 messages belong to 3 categories.

3.3 Manual Annotation

This study summarizes the message analysis and further establishes a classifier of donation intention. The classifier has designed ten category labels, covering the emotional expression and specific intentions of donors, including "Support Again," "Environmental Protection," "Taxation," "Gifts," "Personal Information," "Empathy," "Hope," "Gratitude," "Cheers and Expectations," and "Care and Blessings." These labels reveal their personal motivations, identification with the organization, and expectations for feedback. The distribution and proportion are shown in Table 2. It also enhances the organization and donation connections, fostering a more sustainable and stable support base. To make the manual annotation process more efficient and accurate, the definition of each category needs to be further refined, and examples should be provided. Further refined and examples provided, as shown in Table 1

Since a donation message may contain two or more categories, this study adopts a multi-label tagging method so that messages with multiple meanings can be labeled under different categories.

Label	Annotations	Proportion
Support Again	942	18.0%
Environmental Protection	153	2.9%
Personal Information	316	6.0%
Taxation	228	4.3%
Giveaways	269	5.1%
Gratitude	701	13.4%
Encouragement and Expectations	720	13.7%
Care and Blessings	611	11.7%
Hope	694	13.2%
Empathy	608	11.6%
Total	5242	100%

Table 2. Label Distribution and Proportion

3.4 Experimental Design

In recent years, with the advancement of deep learning, NLP models based on neural networks, especially BERT, have achieved remarkable success in various NLP tasks. This study used the pre-trained BERT model and supervised learning methods to classify donation message data and established a classifier for the donation intention classification problem. To ensure the validity and generalization ability of the model, we used the StratifiedKFold stratified cross-validation method to divide the dataset into five folders with fixed randomness (random state is set to 42). Three approaches were used for model training: deep learning, LLM, and machine learning.

For deep learning, the "bert-base-chinese" pre-trained BERT model was used, and parameters such as learning rate, epochs, and batch size were adjusted to build a binary classification model. Additionally, the GPT-4o-mini LLM model was used for message classification, enhancing classification accuracy through semantic understanding and natural language generation.

In the machine learning part, after tokenization with Jieba, TF-IDF and Word2Vec techniques were applied to transform the text while tuning the regularization parameter (C) and kernel function of SVM, as well as the learning rate, max depth, and the number of decision trees in XGBoost to improve classification accuracy and performance. The research also built an integrated model (VotingClassifier), combining SVM and XGBoost, adjusted the parameters, and used Soft Voting to calculate the prediction probability of each classifier and select the result.

The model performance was evaluated using the Macro-average F1-score, providing a comprehensive analysis of the effectiveness of different methods.

3.5 Evaluation Metric

In this study, the donation intent classification problem belongs to a multi-label task. Therefore, we build ten classifiers for ten donation intents. Each intent category is trained and estimated independently. The Macro-average F1-score measurement metric to evaluate the model's prediction performance of each classifier. The macro-average F1-score calculates the F1-score for each donation intent category and then uses the average of these scores as the performance indicator of the model. This ensures that the ten categories of donation intention are treated as equally important, regardless of the information between categories and the imbalance problem. Macro-average F1-score calculates the F1 score of each category and averages all F1. N is the total number of ten categories. $F1Score_i$ is the F1 score of the i-th category, which is the average precision and recall. The formula is as follows:

$$MacroAverage\ F1Score = \frac{1}{N}\sum_{i=1}^{N} F1Score_i$$

In the results and analysis, comparisons will be made with other models, and Macro-average F1-score will be unified as the final evaluation index.

4 Experimental Results

This research aims to achieve the best combination of performance and efficiency through fine parameter adjustment of deep learning and machine learning models. We conducted extensive performance testing on the BERT model, adjusting parameters such as learning rate and epoch to achieve the optimal configuration. At the same time, the GPT-4o-mini as an LLM model was used for message classification, enhancing classification performance through background definition and semantic understanding. The SVM and XGBoost algorithms were used, and the Kernel type, C parameters, max depth, and number of n_estimators were experimentally adjusted to find the best parameter combination. In addition, the VotingClassifier integrated model was established by combining these two algorithms, and the soft voting method was used to calculate the average value of the predicted probability, which improved the stability and accuracy of the model.

The experimental design of this study significantly improved the prediction accuracy of the model and the robustness of the results, and provided key parameter settings for each model to achieve optimal performance.

4.1 Prediction Performances

For each category in the dataset, the data was divided into five independent folders, each containing training and testing data. On the BERT model, the Macro-average F1-score is 0.9806, and the classification prediction result is 0.8792, which exceeds all test models. The second is the SVM model, which is verified that the F1-score is 0.8731 and the test is 0.8095; the XGBoost model is verified at 0.8408 and the test is 0.7653; while the VotingClassifier model is verified as F1-score is 0.7965, the test is 0.7945; the LLM model performed with a verified F1-score of 0.4989 and a test F1-score of 0.3391. It can be seen that it performs best in BERT classification performance. As shown in Table 3.

Model	Macro-average F1-score	
	Validation Set	Training Set
BERT	0.9806	0.8792
LLM	0.4989	0.3391
SVM	0.8731	0.8095
XGBoost	0.8408	0.7653
Voting Classifier	0.7965	0.7945

Table 3. Classification Effectiveness of Different Classifiers

4.2 BERT Result Analysis

During the training process of the BERT model, we set the learning rate to 1e-5 and the epoch to 13, and obtained the best Macro-average F1-score: 0.9806 on the verification set, as shown in Figure 2. The figure shows that the model with the learning rate set to 1e-5 performs best. As the epoch increases, the F1-score gradually increases and reaches the highest at the 13th epoch, stabilizing and maintaining a high score level. Comprehensive observation, setting the learning rate to 1e-5 can achieve the best results because it achieves a good balance between fast learning and stability of the model.

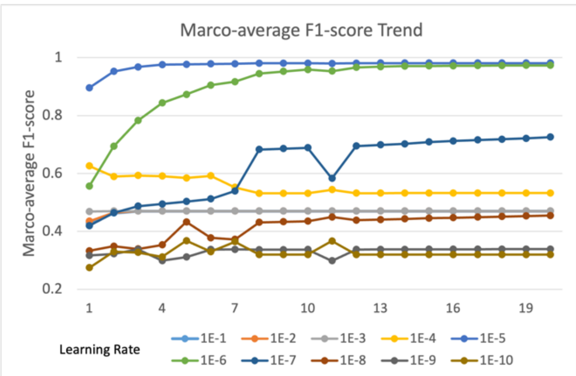


Figure 2. Marco-Average F1-Score Trend

4.3 LLM Result Analysis

When the temperature parameter was adjusted to 0.2, it showed better predictiveness; decisions were more focused and consistent. This change greatly improved the performance and achieved the highest Macro-average F1-score of 0.49 on our validation set, indicating a significant improvement in classification accuracy.

However, these results are still low compared to the others, like BERT, SVM, and XGBoost, demonstrating that LLM is limited in recognizing and classifying donation intents only based on internal knowledge and thinking. Leveraging other techniques such as fine-tuning, multi-task learning, or including external knowledge bases can help the model better classify nuanced and cortical donation intentions.

4.4 SVM Result Analysis

We discovered that the impact of the kernel function on the data set varied greatly when training SVM models using various methods. When using TF-IDF technology for SVM model training, the cost was 10, Kernel was linear, and the best Macro-average F1-score was 0.8731 on the validation set. Figure 3. shows the parameter trend, where the linear kernel acts steadily, and the F1-score rapidly peaks and stabilizes as the cost value rises. On the other hand, the best Macro-average F1-score of 0.4959 was obtained on the validation set when employing Word2Vec technology for SVM model training, with cost set to 10 and Kernel to poly. The performance was relatively low, particularly with the sigmoid and linear kernels, which failed to improve the results significantly.

These findings demonstrate the linear kernel of TF-IDF is more suitable for high-dimensional data with sparse features in text classification, effect-

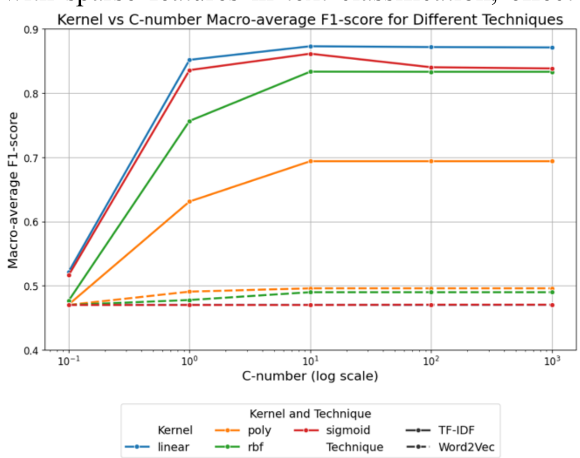


Figure 3. Kernel vs. C-number Macro-average F1-score for Dif-ferent Techniques

tively capturing key semantic features. In comparison, while Word2Vec has the advantage of semantic connection in low-dimensional vector spaces,

its performance is not as good as expected in specific applications, such as multi-category classification.

4.5 XGBoost Result Analysis

During the training of the XGBoost model, when the learning rate was set to 1e-1, maximum depth to 5, and the number of estimators to 300, the best Macro-average F1-score of 0.8408 was achieved on the validation set, as shown in Figure 4. Figure 5 shows that as the learning rate increased, the model performance showed an upward trend but slightly declined after the learning rate reached 0.05. Furthermore, there was a clear interaction between tree depth and learning rate. For instance, at a learning rate of 0.1 and a tree depth of 5, the model reaches the highest F1-score of 0.8408, while at a tree depth of 9, a lower learning rate of 0.01 results in a lower F1-score of 0.82294. As the number of trees increased, model performance became more stable, but excessive depth could lead to overfitting. XGBoost demonstrates strong performance stability in multi-class classification tasks, particularly when handling large datasets and sparse features. Therefore, adjusting the learning rate, tree depth, and the number of trees according to the dataset's characteristics to improve model accuracy and stability across different applications must be carefully balanced to avoid unnecessary computational burden and over-fitting.

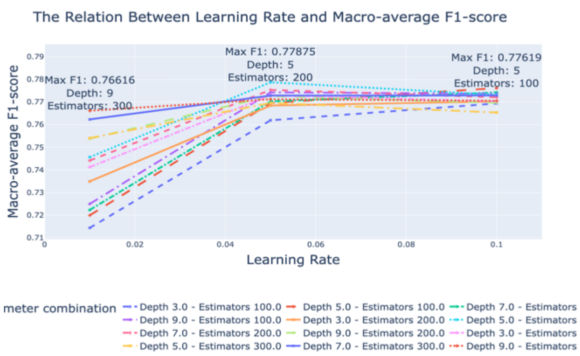


Figure 4. Learning Rate:1e-1 Macro-average F1-score vs Max_Depth for Different N_Estimators

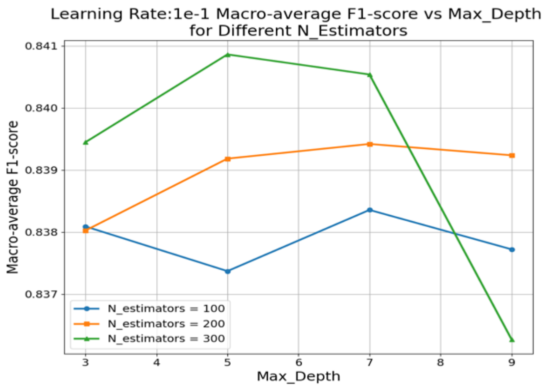


Figure 5. The Relation Between Learning Rate and Macro-average F1-score

4.6 Voting Classifier Result Analysis

The VotingClassifier designed in this section combined the optimal parameter configurations of SVM and XGBoost, using a 'soft' voting strategy to achieve better final classification results. The model achieved a Macro-average F1-score of 0.7945 on the test set, surpassing the performance of the standalone XGBoost, but slightly lower than the F1-score of 0.7965 on the validation set. This shows that although VotingClassifier combines the advantages of both algorithms, its overall performance may still be limited when encountering inconsistencies in handling certain class samples, particularly with edge cases. Theoretically, the integration method can improve the model's generalization ability, but in practical applications, it still requires careful planning and modification to produce better outcomes.

4.7 Misclassification Summary

Table 4 highlights which categories were most affected by model misclassifications. It was found that messages in the "Support and Encouragement" category mostly expressed support and encouragement for recipients, but these emotions are easily related to the "Gratitude" or "Hope" categories. Misclassifications in the “Empathy” category mainly occurred in messages containing abstract or strong emotions, highlighting the limitations of the current models in processing delicate emotions and deep empathy. This study also found that in the “Personal Information” category, the choice of word segmentation tools impacts the model's determination of personal information. For example, the model failed to correctly identify key information in addresses or names, such as the complete address or name in '***Receipt,' leading to unexpected segmentation results and affecting classification accuracy.

Category Labels	Misclassification
Support Again	11
Environmental Protection	5
Personal Information	29
Taxation	9
Giveaways	12
Gratitude	9
Encouragement and Expectations	55
Care and Blessings	43
Hope	15
Empathy	56
Total	244

Table 4. Statistics of Models' Misclassifications

5 Conclusions

In conclusion, through detailed analysis and improvement, this study aims to introduce deeper emotion recognition techniques or strengthen the semantic understanding ability of the model in future applications, helping the model to more effectively process texts involving complex emotions and abstract concepts, thereby improving overall classification performance.

Acknowledgments

This research was partially supported by the Ministry of Science and Technology, Taiwan (grant number: MOST 111-2221-E-031-004 -MY3) and the Center for Applied Artificial Intelligence Research of Soochow University.

References

Cheng-Xun Hsiao. 2020. A Study on People's Behavior and Intention of Donation: Case of Donors of a Residential Home for Disability in Chiayi City. Master's thesis. National Chi Nan University, Puli, Taiwan. Advisor: Yuan-Hsieh Huang.

Chia-Chun Sun. 2023. An Analysis of Deliberative Democracy and Political Polarization in Social Media Dialogue: A Case of YouTube Live Chat Messages on CDC COVID-19 Pandemic Press Conference. Master's thesis. National Taiwan University, Taipei, Taiwan. Advisor: Ji-Lung Hsieh.

Chih-Yi Lin. 2020. Integrating Sentiment Analysis and Text Mining for News-headlines to Predict FOREX Market. Master's thesis. National Taipei University of Technology, Taipei, Taiwan. Advisor: Chen Shu Wang.

Chueh-Tu Yang. 2023. Egoistic or Altruistic? - How Personality Impacts the Affection of Prosocial Referral Behavior through Motivation. Ph.D. dissertation. Tamkang University, New Taipei City, Taiwan. Advisor: Pei-Chi Lee.

Hsiu-Wen Tseng. 2023. Predicting Fundraising Performance Using Image Framing and Message Framing: A Case of Apple Foundation. Master's thesis. Department of Industrial Engineering and Management, National Chiao Tung University, Hsinchu, Taiwan. Advisor: Dr. Sih-Han Chen.

Marudas, N. P., T. Hahn, and F. A. Jacobs. 2012. "An Improved Model of Donations to Nonprofit Organizations." In Proceedings of ASBBS Annual Conference, vol. 19, no. 1, 545-559. Las Vegas.

Miao-Hsien Hsueh. 2014. A Study of Relationship Marketing of NPO for Fundraising: A Case of Noordhoff Craniofacial Foundation. Master's thesis. Chaoyang University of Technology, Taichung, Taiwan. Advisor: Ji-Lung Hsieh.

Min-Chou Lee. 2023. Strategic Analysis of Ecosystem Establishment by Instant Messaging Software: A Case Study of LINE and its Development in Taiwan. Master's Thesis. Graduate Institute of Business Administration, College of Management, National Taiwan University, Taipei, Taiwan. Advisor: Ruey-Shan Guo.

OpenAI. 2023. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

Ru Huang and Xiuli Wang. 2023. Impact of COVID-19 on Mental Health in China: Analysis Based on Sentiment Knowledge Enhanced Pre-training and XGBoost Algorithm. Frontiers in Public Health, 11, Article 1170838.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 785–794.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In Findings of EMNLP.

Yang, Kung-Lin. 2022. Forecasting Taiwan Index Futures With Machine Learning Based - A Predictive Study. Master's thesis. Soochow University.

Yih-Hui Lu and Ching-Shou Lin. 2013. A Study on the Behavioral Intention of Charitable Donation for Middle-Aged and Senior Citizens. Journal of Welfare Technology and Service Management 1, 3 (2013), 35-50.

Yu-Chieh Wang. 2023. Analysis of Donation Expenditure Before and After the Pandemic. Master's thesis. National Central University, Taoyuan, Taiwan.

Yu-Ting Kuang. 2015. Message Analysis and User Classification in Facebook Pages. Master's Thesis. Department of Information Management, National Taiwan University of Science and Technology, Taipei, Taiwan. Advisor: Yung-Ho Lu.

Wan-Ling Chen. 2014. The Impact of International Experiences and Personality Traits on Donation Willingness to Non-profit Organizations. Master's thesis. Chang Jung Christian University, Tainan, Taiwan. Advisor: Yi-Hui Ho.

Wei-Hong Lin. 2020. Implementing a Food Rating System Using Instagram Data. Master's Thesis. Department of Information Management, College of Management, National Taiwan University, Taipei, Taiwan. Advisor: Yuh-Jzer Joung.

Assessing the Necessity and Impact of Localized Traditional Chinese Function Calling Benchmarks

Liang-Chieh Lee[†] **Cheng-Wei Lin**[†] **Pei-Chen Ho**[†] **Da-shan Shiu**

MediaTek Research
[†]Internship.

Abstract

The function-calling capability of Large Language Models (LLMs) is becoming indispensable for their practical applications. For LLMs to be successfully applied to localized commercial use, function calling refers to the ability to invoke external tools to obtain real-time information or interact with additional functionalities. To develop or select the ideal models for these tasks, it is crucial to understand the importance of benchmark localization.

In this study, we introduce our recreation of a Taiwan-specific standardized function-calling benchmark, adapted from the Gorilla function-calling framework for evaluating tool calls in English. Through experimental evaluation utilizing our formed data, question-answer scoring mechanisms, and additional tools for multilingual performance comparison, we successfully completed the zh-TW localization process and assessed its differences compared to the English evaluation. This highlights the necessity of evaluating local Traditional Chinese performance, as it provides a clearer perspective on localized applications in commercial contexts and other fields in Taiwan.

Keywords: LLM function calling, Traditional Chinese

1 Introduction

While large language models (LLMs) have shown remarkable abilities in generating text and reasoning, relying solely on the model's internal capabilities presents certain limitations. Traditional LLMs like GPT-3 depend on static reasoning, which restricts their ability to update information in real-time(Yao et al., 2023). This often leads to fact hallucinations and error propagation, especially in multi-step reasoning processes. Furthermore, LLMs struggle with integrating external knowledge, as they are limited to the information embedded in their training data. Without the ability to retrieve and incorporate real-time, domain-specific information from external environments, LLMs can generate inaccurate or outdated responses, making them unreliable for many real-world applications.

Function calling (or tool calling) provides a robust solution to these limitations by allowing LLMs to access external tools and APIs. This enables models to retrieve up-to-date information and perform tasks that require specialized knowledge or computations, such as complex mathematical calculations or accessing real-time data from databases. By reducing reliance on static internal reasoning and outsourcing specific tasks to reliable external sources, function calling helps minimize hallucinations and enhances task accuracy(Schick et al., 2023; Abdelaziz et al., 2024). In commercial applications, this capability is crucial as it allows businesses to leverage LLMs for dynamic tasks such as financial analysis, customer service automation, and data retrieval, ensuring the responses are accurate, timely, and grounded in the latest information. To more accurately understand the potential business applications within the Taiwan region, we aim to provide a more precise description of how such function calls perform.

2 Related Works

Due to the significant differences in the availability of training data between high-resource and low-resource languages, we suspect that the function calling capabilities of Traditional Chinese may differ from English(Hsu et al., 2024). For example, English occupies 89.7% of the pretraining corpus, far surpassing the mere 0.13% for Chinese(Zhu et al., 2024; Li et al., 2024). High-resource languages typically perform more stably in large language models (LLMs) due to the support of extensive datasets, while low-resource languages like

Traditional Chinese may experience performance degradation due to insufficient data(Lin and Chen, 2023). Specifically, there are notable differences in the syntactic structures of Traditional Chinese and English. For instance, the subject-verb-object structure in Chinese is more flexible, and the placement of verbs is less fixed compared to English. Moreover, Chinese employs phrases in a distinct manner from English, and these features could potentially affect the function calling performance of LLMs in Traditional Chinese(Chang et al., 2024). Therefore, we have reason to believe that these syntactic differences may influence the performance of LLMs in both languages(Nezhad and Agrawal, 2024).

Based on this assumption, we aim to conduct a comprehensive evaluation of function calling in Traditional Chinese, using a series of tests to analyze whether language-specific characteristics affect performance during function calls. These tests will help us clarify the specific performance of LLMs in Traditional Chinese contexts and determine if there are areas that require further adjustments. By doing so, we can more accurately assess the usability of Traditional Chinese in these scenarios. This approach not only improves the accuracy of function calling in Traditional Chinese but also verifies whether linguistic differences play a crucial role in these technical applications.

3 Methodology

To validate the need for localized large language models (LLMs) in Taiwan using Traditional Chinese for function calling, we adopted the benchmarking methodology from Gorilla's APIBench(Patil et al., 2023). We improved the original dataset, which was not well-suited for Traditional Chinese, and refined the evaluation criteria for return values. Additionally, we retained key strengths of the framework, such as automated API generation and invocation. APIBench offers one of the most comprehensive API datasets in the machine learning field, significantly reducing data errors and risks, and providing a solid foundation for handling Traditional Chinese in subsequent processes.

Upon completing the data operations, we implemented a language configuration feature and conducted several bilingual evaluations of well-known, large-scale LLMs. This allowed us to assess whether performance differences exist between languages and to compare their effectiveness in different linguistic environments.

3.1 Dataset Configuration

The dataset collection for APIBench comes from recording model cards on the three major platforms: HuggingFace, PyTorch Hub, and TensorFlow Hub. Incomplete models were filtered out, resulting in a total of 1,645 API calls. These model cards were then converted into JSON format, and GPT-4 was used to generate synthetic instruction data, creating 10 instruction-API pairs for each model(Patil et al., 2023).

The dataset utilized in the Gorilla experiment provides a comprehensive analysis across various user applications, including its use in proxies and enterprise workflows. This dataset, encompassing a wide range of topics and fields, can be able to hold equivalent evaluative value in the Traditional Chinese context.

The evaluation metrics are categorized into Python and non-Python, and corresponding Traditional Chinese datasets are established.

Python evaluations include:
- **Simple function:** Evaluates a single function call using a JSON function document.
- **Multiple function:** Requires the model to select the best function to invoke from 2 to 4 JSON function documents.
- **Parallel function:** Involves invoking multiple function calls in parallel for a single user query.
- **Parallel multiple functions:** Combines parallel and multiple functions, where multiple function documents are provided, and each corresponding function call is invoked zero or more times.

Each category is assessed using both Abstract Syntax Tree (AST) and executable function evaluations.

Non-Python evaluations include:
- **Function relevance detection:** Checks if the model correctly identifies when no provided functions are relevant.
- **REST API:** Tests the model's ability to generate executable REST API calls using real-world GET requests, including path parameters and key/value pairs.
- **SQL:** Assesses the model's capability to construct reliable SQL queries using customized functions.
- **Java and JavaScript:** Tests the model's ability

to handle language-specific types, such as Java's 'HashMap'.

We conducted a detailed review of the question-answer pairs in these datasets, selecting translations that feature localized Taiwanese terminology for queries, adjusting both syntax and content. In addition to invoking external APIs for accurate evaluation, we mapped the results to the standard answers and incorporated key Chinese keywords. This manual approach ensures that the translated content better reflects the authentic usage of Traditional Chinese syntax. After translating the question-answer pairs, we designed a systematic distribution configuration, allowing language-based assignments to be tested with corresponding datasets. This recreation provides a more convenient and intuitive method for evaluating the comparability of large language models across different languages.

3.2 Benchmarking Framework

In our study, we first utilized the Abstract Syntax Tree (AST) as a core tool for program compilation and parsing. AST represents the syntax of a program in a tree structure, breaking down syntactic elements into various nodes, with each node representing a fundamental unit of the program's syntax. By stripping away the syntactic details and preserving only the semantic structure of the code, AST aids in more efficient program compilation, optimization, and analysis. In the context of API call validation, AST is employed to parse the API calls generated by the model, progressively examining the syntactic structure to ensure consistency with reference documentation. AST also plays a critical role in handling parameter types, nested structures across different languages, and identifying model hallucinations, which refer to API calls that do not match any known API in the database. Moreover, AST proves useful in validating multiple and parallel function calls by efficiently parsing and checking the syntactic structure of each function to ensure the accuracy of API calls.

Following this, we introduced Executable Function Evaluation, which validates the correctness of generated API calls by executing them. This method is divided into non-REST and REST types. In non-REST evaluation, the output is assessed based on three criteria: exact match, real-time match, and structural match. REST evaluation, on the other hand, focuses on the successful execution of API calls and ensures the type and structure of JSON responses are consistent. Given that REST responses may vary over time, the evaluation emphasizes structural consistency rather than static values. The multiple and parallel function evaluations extend the principles of single-function evaluations by comparing the model-generated outputs with ground truth values to ensure that all outputs meet the evaluation criteria.

3.3 Evaluation

In the APIBench framework, test data are provided in the form of test files across all evaluation categories (see Table 1), significantly reducing the effort required to reformat various types of responses. Leveraging the language configuration feature developed, we conducted function-calling benchmark evaluations in both Traditional Chinese and English. By employing parallel question-answer pairs, we aim to assess the models' performance and function-calling capabilities when posed with questions in both languages. To ensure robustness and generalizability of the results, we evaluated several widely recognized models, including GPT-3.5, GPT-4o, Claude 3.5, and Gorilla OpenFunctions, which was trained using APIBench results.

4 Result and Discussion

To enhance the comparability of cross-linguistic function-calling benchmarks, we executed a bilingual comparison script and created radar charts based on the most critical evaluation metrics: simple function calls, multiple function calls, parallel function calls, parallel multiple function calls, executable simple calls, executable multiple calls, executable parallel calls, executable parallel multiple calls, and relevance detection. This functionality was integrated into our new configuration, along with a language-switching feature, to provide a reliable evaluation across multiple languages. Through a comparative evaluation of model performance in Traditional Chinese (see Figure 1) and English (see Figure 2), we found a noticeable gap in performance when models were tested in Traditional Chinese compared to English. This observation supports our previous hypothesis: the function-calling capabilities are influenced by linguistic differences, particularly in the case of Traditional Chinese or the language culture used in the Taiwan region.

Model	IR	AST				EXEC			
		S.	M.	P.	P. M.	S.	M.	P.	P. M.
gpt_35_turbo_0125 (FC)	7.5	70.0	73.5	67.0	49.5	62.4	90.0	76.0	52.5
gpt_4o_2024_05_13 (FC)	64.2	67.7	73.0	76.0	58.5	64.7	84.0	80.0	70.0
gorilla_openfunctions_v2 (FC)	52.1	66.7	65.5	59.0	43.5	61.2	92.0	62.0	57.5
claude-3.5-sonnet (FC)	82.1	74.9	79.0	77.0	67.5	66.0	94.0	86.0	65.0

Table 1: **Function calling benchmark.** This table shows the accuracy across the four models with the ability of function calling. **IR** denotes "irrelevance detection". **AST** denotes "abstract syntax tree". **EXEC** denotes "execution". **S.** denotes the case of simple function. **M.** denotes the case of multiple function. **P.** denotes the case of parallel function. **P. M.** denotes the case of parallel multiple function.

Figure 1: **Comparison between our Traditional Chinese benchmark and Berkeley function calling leaderboard.**

4.1 Overall Performance

Our study evaluated GPT-3.5, GPT-4, and Gorilla OpenFunction, and we found that the overall performance in Chinese was slightly lower compared to English. Additionally, the performance trends between Traditional Chinese and English across different categories displayed a discernible pattern. However, this general trend is not uniform across all categories, highlighting the complexity of language model performance across different tasks and languages. A notable exception to the overall trend is observed in the evaluation of executable multiple function calls. In this specific category, the performance does not adhere to the observed pattern of English outperforming Traditional Chinese. For instance, in the Multiple (Exec) category, the Chinese version of GPT-4 (84%) outperforms its English counterpart (78%). Similarly, the Chinese version of Gorilla OpenFunctions v2 achieves 92% accuracy in this category, compared to 94% for its English version, reflecting a much smaller gap than in other categories.

4.2 Implications for Traditional Chinese LLMs

The results highlight several key points for the development and application of LLMs in Traditional Chinese: Language-specific fine-tuning: The performance gap between English and Traditional Chinese suggests a need for more extensive and targeted fine-tuning for Traditional Chinese models. Task complexity: As task complexity increases, the performance in Traditional Chinese tends to degrade more rapidly than in English. This indicates a need for more diverse and complex Traditional Chinese datasets for training. Model architecture: The varied performance across different models suggests that certain architectures may be more suitable for handling Traditional Chinese function calls. Further research into model architectures optimized for Traditional Chinese could yield significant improvements. Data quality and quantity: The generally lower performance in Traditional Chi-

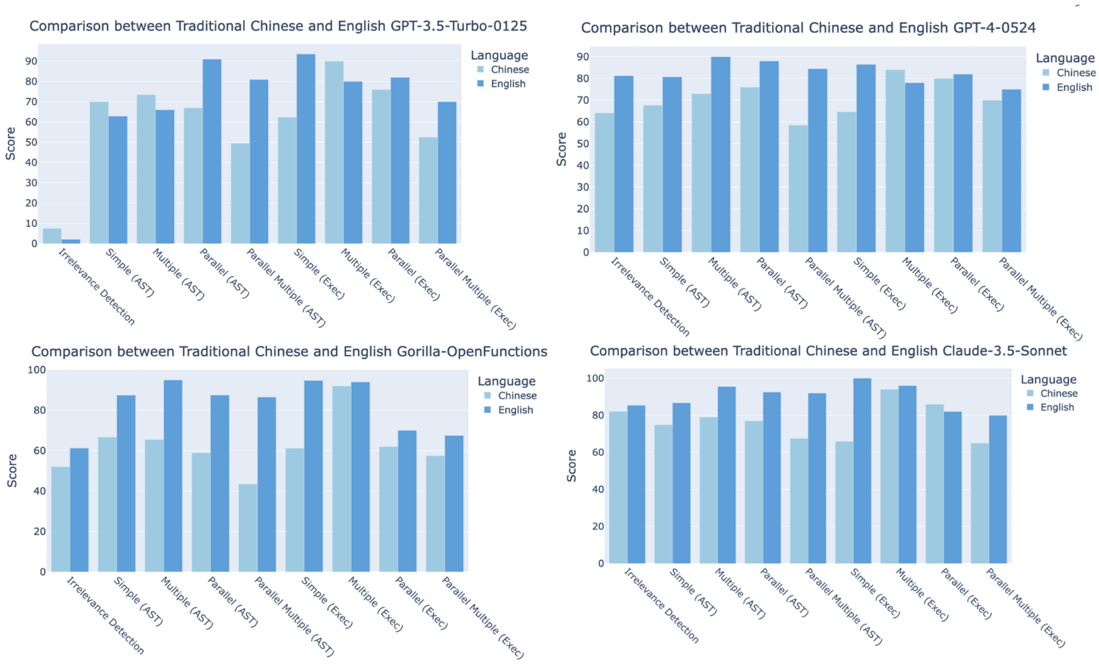


Figure 2: **Performance comparison between Traditional Chinese and English.**

nese across most categories underscores the need for larger, high quality Traditional Chinese datasets, particularly for complex function calling tasks. Cross-lingual transfer: The performance discrepancies between languages suggest that cross-lingual transfer learning techniques could be explored to leverage the strengths of English models in improving Traditional Chinese models. Task-specific optimization: The inconsistent performance across different categories highlights the importance of task-specific optimization. Rather than applying a one-size-fits-all approach to improving Traditional Chinese LLMs, developers should consider tailoring their approaches based on the specific types of function calls and tasks that are most critical for their applications.

5 Conclusion

The results of this experiment clearly demonstrate that there are significant differences in the performance of function-calling benchmarks between Traditional Chinese and English. It is evident that localized evaluations are crucial in contexts where Traditional Chinese is used exclusively. Due to the significant syntactic differences between Traditional Chinese and English, such as the lack of fixed part-of-speech positioning and the distinct structure of phrases, these linguistic disparities are reflected in function-calling performance. Therefore, evaluation methodologies should be developed somewhat independently from those used for English models. In the context of function-calling applications in Taiwan, the importance of Traditional Chinese function calls cannot be underestimated. Language-specific fine-tuning is essential for achieving localized commercial applications. Although English models generally outperform Traditional Chinese models, there are certain tasks where the performance of Traditional Chinese models exceeds that of their English counterparts. This suggests that, when considering the practical application of large language models, local developers should carefully consider the appropriate contexts and timing for Traditional Chinese function-calling models, selecting the models that best suit the tasks at hand. This approach will further maximize the practicality and reliability of Traditional Chinese models across various applications.

6 References

References

Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Sadhana Kumaravel, Matthew Stallone, Rameswar Panda, Yara Rizk, GP Bhargav, Maxwell Crouse, Chulaka Gunasekara, et al. 2024. Granite-function calling model: Introducing function calling abilities via multi-task learning of granular tasks. *arXiv preprint arXiv:2407.00121*.

Ting-Yun Chang, Jesse Thomason, and Robin Jia. 2024.

Do localization methods actually localize memorized data in llms? a tale of two benchmarks. *arXiv preprint arXiv:2311.09060*.

Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da-Shan Shiu. 2024. Breeze-7b technical report. *arXiv preprint arXiv:2403.02712*.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages. *arXiv preprint arXiv:2404.11553*.

Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*.

Sina Bagheri Nezhad and Ameeta Agrawal. 2024. What drives performance in multilingual language models? *arXiv preprint arXiv:2404.19159*.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *ArXiv preprint arXiv:2302.04761*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Lingxuan Zhu, Weiming Mou, Yancheng Lai, Junda Lin, and Peng Luo. 2024. Language and cultural bias in ai: comparing the performance of large language models developed in different countries on traditional chinese medicine highlights the need for localized models. *Journal of Translational Medicine*, 22(1):319.

Beyond Fine-Tuning: A Non-Parametric Approach to Distractor Synthesis for Multiple-Choice Questions

Yu-Chen Cheng
National Chung Hsing University

Yao-Chung Fan*
National Chung Hsing University yfan@nchu.edu.tw

Abstract

Automated distractor generation is crucial for creating effective multiple-choice questions. Traditional methods often require fine-tuning language models with domain-specific data, limiting adaptability and scope control. This paper introduces a new, non-parametric framework that uses machine reading comprehension to generate contextually relevant yet incorrect distractors from hard negative passages, without the need for fine-tuning. This approach allows rapid deployment across various domains and enables educators to tailor questions to specific content. Our framework outperformed state-of-the-art models by 8 percentage points on in-domain datasets and 75 percentage points on out-domain datasets, offering greater adaptability and controllability, making it more suitable for educational use.

Keywords: Multiple-Choice Questions, Distractor Generation, Non-parametric Framework

1 Introduction

Effective distractor selection are crucial in automated assessment systems for evaluating the depth of a learner's understanding. Creating such distractors traditionally requires expert human, making it a significant challenge in automated systems. Therefore, there has been an increasing focus on automating distractor generation through research (Liang et al., 2017, 2018; Chiang et al., 2022; Wang et al., 2023).

The common setting of multiple choice question (MCQ) is as shown in Figure 1. For automatic generation, the setting of distractor generation (DG) task is to take (1) a question stem Q and (2) the corresponding answer A as input. The goal of the output is to have a set of distractors (should be relevant but wrong with respect to Q and A).

Question ：
Which organ is responsible for the removal of excess water, salts, and waste products from the blood?
Answer ：
Kidneys
Distractors:
A) Heart B) Lungs C) Liver

Figure 1: A MCQ Test Example: the challenge to MCQ test preparation lies in wrong option (distractor) selection.

The existing DG methods, as discussed in recent studies (Wang et al., 2023; Ren and Q. Zhu, 2021; Chung et al., 2020), involve fine-tuning language models (LMs) using specialized DG datasets (such as CLOTH (Xie et al., 2017) or MCQ (Ren and Q. Zhu, 2021)), as illustrated in the left part of Figure 2. Despite the progress in current DG methodologies, there is still significant room for enhancement. First, fine-tuning approaches require domain-specific adaptation when transitioning to new fields. For instance, a DG model trained for the medical field cannot be directly applied to generate content for the scientific domain. Second, certain educational scenarios demand control over the question scope, such as restricting the generated questions and distractors to a specific range or the currently taught scope.

Addressing these limitations, we propose a non-parametric framework that begins by retrieving relevant contexts and then *extracting wrong answers* from these contexts. At the heart of our framework is the use of an *Extractive Reader*. In the context of machine reading comprehension (MRC) (Zhang et al., 2021), an

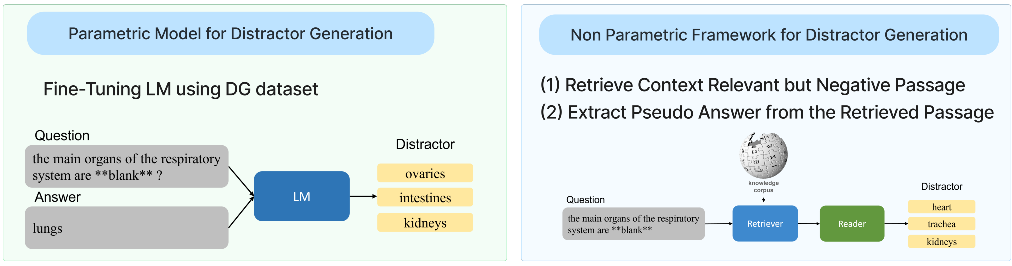


Figure 2: Contextual Distractor Synthesizer: This figure contrasts our proposed non-parametric framework (the left part of the figure) for distractor generation (right) with the existing fine-tuned Language Models (LM) approach (left) (Wang et al., 2023; Ren and Q. Zhu, 2021; Chung et al., 2020). Our framework operates by first retrieving contextually relevant passages that do not contain the correct answers. It then extracts pseudo answers from these passages with a machine reading comprehension model, utilizing them as potential distractor candidates.

Extractive Reader is a type of model designed to answer questions by identifying and extracting spans of text directly from a given document. This approach assumes that the answer to any question is a span of the document's text—typically a phrase or a sentence.

We utilized an *Extractive Reader* as the *Contextual Distractor Synthesizer*. Pre-selected hard negative passages and question stems were input into the *Contextual Distractor Synthesizer*, enabling us to generate pseudo answers—plausible but incorrect distractors.

Distractor Synthesis Process: Upon receiving a relevant passage (which does not contain the correct answer) and a question stem, the Contextual Distractor Synthesizer generates pseudo answers that are contextually relevant to the question stem, using the passage's content to ensure their plausibility as distractors. These pseudo answers are incorrect by design, serving as high-quality distractors.

The use of hard negative passages in our methodology ensures a balance between relevance and incorrectness, creating distractors that are both challenging and convincing. This approach significantly increases the cognitive demands on learners, thereby improving the quality of multiple-choice assessments.

The following features distinguish our DG method:

- **No Need for Fine-Tuning:** Setting our method apart from other DG models that require extensive fine-tuning, our approach operates without the need for specific training. This attribute allows for swift adaptability across various domains, as changing the knowledge corpus is sufficient to tailor distractor generation to different subject areas.

- **Control over Question Scope:** Our framework offers the unique ability to control the thematic scope of the questions. By exchanging the data corpora, we can selectively focus on generating questions within specific topic boundaries. This aspect of our methodology resonates with the practical necessities of educational contexts, where educators often seek to limit question topics to relevant subject matter. This level of control over the question scope is a feature not readily achievable in Text2Text models(Wang et al., 2023).

2 Related Work

The landscape of distractor generation research is currently delineated into two primary frameworks, each with its distinct methodologies and advancements.

Generating and Ranking (GR) Framework: This framework employs a general-purpose knowledge base to create a candidate set of distractors, followed by a feature-rich learning-to-rank model for distractor selection. The GR architecture operates in two stages: initial generation of candidate distractors and subsequent ranking based on semantic rules and linguistic features. There are two prevalent approaches within this framework: one utilizes a knowledge base (Ren and Zhu, 2021),

	Method Type		Adaptability	Controlability	Model
	Extractive	Generative			Type
Gao et al. 2019		Y			RNN
Zhou et al. 2020		Y			RNN
Araki et al. 2016	Y		Y		Non-neural model
Welbl et al. 2017	Y				Random forests
Guo et al. 2016	Y				Word2Vec
Kumar et al. 2015	Y		Y		SVM
Liang et al. 2017		Y			GAN
Liang et al. 2018	Y		Y		neural/feature-based model
Chung et al. 2020		Y			PLM
Ren and Q. Zhu 2021		Y	Y		Knowledge-base
Peng et al. 2022		Y			PLM
Chiang et al., 2022		Y			PLM
Wang et al., 2023		Y			Text2Text
Our work	Y		Y	Y	Retriever-Reader

Table 1: An Overview of the Existing Distractor Generation Methods: Adaptability: the ability to swift over various domains. Controlability: the ability to control over generated question scope

and the other leverages a language model (Chiang et al., 2022). These approaches have marked a significant improvement over traditional rule-based methods (Liang et al., 2017, 2018), offering enhanced quality and diversity in distractors and are considered state-of-the-art in DG.

Text2Text Generation Architecture: The Text2Text generation framework formulates distractor generation as a Text2Text task, diverging from the GR approach. It involves concatenating the question stem with the correct answer and feeding this combined input into a generative language model (e.g., T5 or GPT). This architecture trains the model to generate a set of distractors in a direct and streamlined manner. Recent research adopting the Text2Text model (Wang et al., 2023) has set new benchmarks in distractor generation, demonstrating state-of-the-art performance.

Our Contribution - Contextual Distractor Synthesizer: Our work introduces a novel paradigm in the field of DG, distinct from both the GR and Text2Text architectures. At the core of our methodology is the use of a Machine Reading Comprehension (MRC) Reader, adapted as a *Contextual Distractor Synthesizer*. Unlike the GR framework, which relies on knowledge bases or LMs for initial distractor generation, our approach utilizes pre-selected hard negative passages that are contextually aligned but factually divergent from the correct answer. This method allows for the generation of pseudo answers that are inherently incorrect but contextually relevant, enhancing the cognitive challenge in assessments.

Another significant feature is that in the real world, teachers often need to create exam questions within a specified range of topics. Previous research has not considered that generated distractors need to be within a designated scope, which can cause differences between experimental results and real-world applications. Our approach is corpus-dependent, ensuring that by providing a specified corpus, the generated distractors will be 100% within the exam scope. This method more closely aligns with practical use cases.

Moreover, our approach does not require the training associated with the Text2Text models. By swapping out the knowledge corpus, our method easily adapts to various domains, offering flexibility and control over the scope of questions. This feature is particularly advantageous in educational settings where specificity and relevance are the key.

For clarity of comparison, we summarize the existing DG studies in Table 1. Our study is the only one capable of adapting to domain changes without requiring any processing or retraining, while also customizing the scope of questions for the same problem.

In Figure 5, although our performance on the in-domain dataset was comparable to that of the current SOTA DG method, we significantly excelled in the out-domain dataset.

This underscores the limited adaptability of the SOTA DG method when applied beyond its initial domain.

3 Methodology

3.1 Problem Setting and Assumptions

The methodology assumes the following inputs :

- A knowledge corpus $\mathcal{C}$, consisting of text chunks relevant to the subject matter.
- A question stem Q, representing the query to be addressed.
- An answer A, the correct response to the question stem Q.

Additionally, the methodology utilizes two functions:

- A document retriever $\mathcal{F}()$, which fetches text chunks from the corpus $\mathcal{C}$.
- An extractive document reader $\mathcal{R}(Q)$, which extracts answers from a fetched text chunk given the question stem Q.

The details of the algorithm are presented in Algorithm 1.

3.2 Retriever

The first stage of the process is the retrieval of relevant text chunks, a critical step in identifying suitable passages for distractor generation. We use the BM25 algorithm to retrieve relevant text chunks for distractor generation.

The passages containing the correct answer and synonyms of the correct answer are filtered out, and the remaining passages are referred to as hard negative passages. This ensures contextually relevant yet misleading content for effective distractor generation.

3.3 Extractive Reader as Contextual Distractor Synthesizer

At this stage, a Extractive Reader is employed as a Contextual Distractor Synthesizer (CDS). This approach deviates from traditional reading comprehension question-answering methodologies by shifting focus from extracting accurate answers to generating contextually relevant yet incorrect distractors, termed as *pseudo answer* (PA) as served as candidates for distractors. The *pseudoanswer* are generated using the following formula:

$$P_{CDS}(PA \mid C, Q) = \prod_{i=1}^{n} P(pa_i \mid pa_1, \ldots, pa_{i-1}, C, Q; \theta)$$

3.4 Formation of the Pseudo Answer Set (PAS):

PAS, or the Pseudo Answer Set, is a collection of all potential distractors generated by the MRC Reader. It is formulated based on two key parameters:

- k —the number of text chunks retrieved by the Retriever
- h —the number of pseudo answers identified by the Reader for each text chunk.

Thus, PAS comprises a total of $k \times h$ pseudo answers, each representing a potential distractor candidiate derived from the hard negative passages.

Character-Level Rouge Score Evaluation: To enhance the quality and diversity of PAS, each pseudo answer undergoes a character-level Rouge score evaluation before adding into PAS by Considering each individual character as a complete word. This evaluation assesses textual similarity and ensures that the pseudo answers are distinct from one another. Specifically, If a pseudo answer candidate (pa) has a high Rouge score compared to any existing item in PAS ($PA_{\in PAS}$), indicating a significant overlap, it is discarded to prevent redundancy.

3.5 Distractor Evaluator

The culminating phase of our distractor generation process is the evaluation of the pseudo answers by the Distractor Evaluator. This component is critical for appraising the $k \times h$ pseudo answers produced by the Contextual Distractor Synthesizer, and its primary goal is to determine the most appropriate distractors for each question based on their relevance and plausibility.

Scores of Pseudo Answers: The input to the Distractor Evaluator comprises the question stem Q, the hard negative article C, the correct answer A, and the set of pseudo answers $PAS = \{PA_i\}$. The pseudo answers are

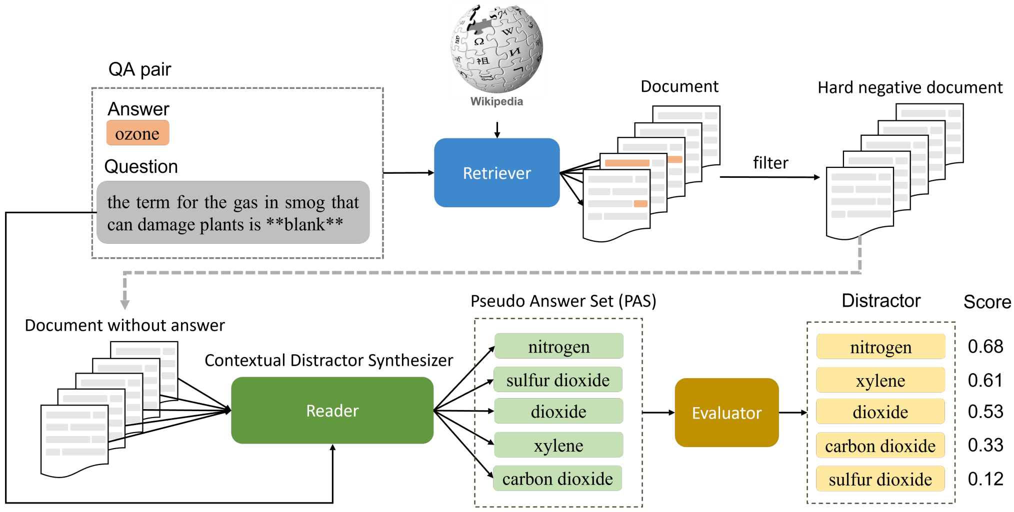


Figure 3: Overview of our approach. We can decompose our architecture into the following components: Retriever, Contextual Distractor Synthesizer, Distractor Evaluator. Initially, the retriever component searches for relevant articles and filters out those containing the correct answers, leaving only the hard negative passages. The Contextual Distractor Synthesizer component then generates pseudo answers from these hard negative passages. Finally, the evaluator component ranks the top-n distractors.

ranked based on the Confidence Score given by the following features:

- **Retriever Score S_{retr}:**

$$S_{retr} = retr(Q, C)$$

  This score, represents the relevence between Q and C. Any value that can represent the relevance between Q and C in sparse retrieval or dense retrieval can be used.

- **Confidence Score $S_{confidence}$:**

$$S_{confidence} = p_{CDS}(PA|Q, C; \theta)$$

  This score, reflects the confidence of the Contextual Distractor Synthesizer (CDS) in PA being a viable distractor.

Selection of Distractors: Each pseudo answer PA receives a final score based on its S_{retr} and $S_{confidence}$ scores:

$$score(PA_i) = S_{retr} \cdot S_{confidence}$$

The distractors that achieve the highest scores in this evaluation —the top-n scored items —are selected as the final distractors. This process ensures that the chosen distractors are both contextually relevant and sufficiently challenging.

4 Experiment

5 Implementation Details

In our experiments with the MCQ dataset and MEDMCQA, we utilized Wikipedia articles as the corpus for our retriever. Each article was divided into passages every five sentences, with the article title appended to the beginning of each passage. We used Pyserini to construct our BM25 retriever model, effectively identifying relevant passages from the corpus.

For the distractor synthesizer component of our framework, we selected Llama-2-7b-hf, released by Facebook Meta on Huggingface, and trained it over 2 epochs using two NVIDIA RTX 3090 GPUs with the SQuAD dataset (Rajpurkar et al., 2016). We used the question answering task to train the model. After finetuning the model, we can utilize relevant text chunks and questions as model inputs to generate pseudo answers. All experiments are conducted using two NVIDIA RTX GPUs.

5.1 Evaluation Metrics

We introduce the GPT-4 Distractor Effectiveness Index (GDEI) to overcome the limitations of traditional token-based metrics in evaluating distractor quality. Unlike token scores, which often miss semantic and contextual details crucial for effective distractors and are constrained by the dataset's limited ground

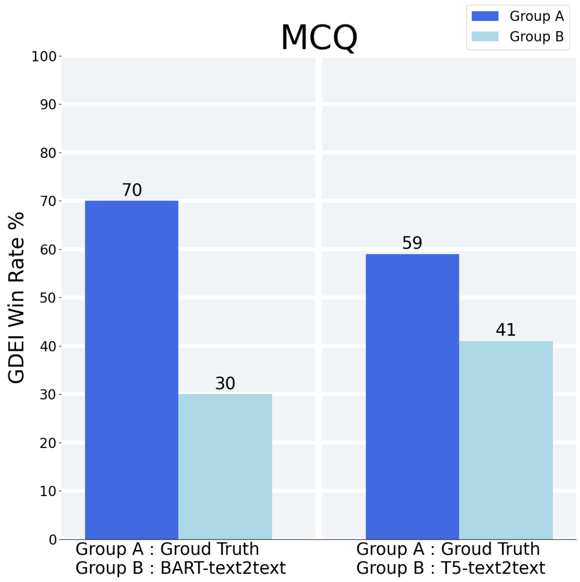


Figure 4: We aim to demonstrate the capability of the GDEI metric. Given that DG-text2text(Wang et al., 2023) which is the SOTA model employs Multiple Choice Questions (MCQ) for training, we tasked GPT-4 with comparing between the ground truth in MCQ and distractors generated by the DG-text2text. The results indicate that GPT-4 is capable of correctly differentiate instances where distractors generate by DG-text2text are lower in quality than the ground truth.

truth, GDEI utilizes GPT-4' s advanced comprehension to assess distractor sets more holistically.

In the evaluation process, GPT-4 is given a question, the correct answer, and two sets of distractors. It then determines which set is of higher quality. To ensure fairness, the order of the distractor sets is randomly alternated. Detailed instructions are provided in Figure 10.

5.2 GDEI Validation through Comparative Analysis

Our validation contrasts GDEI scores for educator-crafted ground truth distractors against those generated by T5 and BART models. The results in Figure 4 favor ground truth distractors, showcasing GDEI's capability to discern quality reflecting human expertise in distractor design. This confirms GDEI's effectiveness as a nuanced evaluation tool for distractor generation.

5.3 Dataset

- **MCQ dataset**: MCQ dataset (Ren and Q. Zhu, 2021) is a cloze-style dataset, that includes the domains of science, vocabulary, common sense, and trivia. Each data is composed of a sentence containing " **blank** " of cloze stem, answer, and distractors.

- **MedmcQA dataset**: MedMCQA (Pal et al., 2022) MedMCQA is a vast dataset of over 194k high-quality Multiple-Choice Questions and Answers for medical entrance exam preparation. It covers 2.4k healthcare topics and 21 medical subjects from AIIMS and NEET PG exams. The questions vary in length and complexity, and each sample includes a question, correct answer(s), additional options, and a detailed solution explanation.

- **Harry Potter Novel questions**: The data from the Harry Potter novels can serve as an excellent dataset for testing Controllability of the generation method. Each book in the series contains unique plot elements not found in the others. We utilized ChatGPT to generate 100 sets of questions from the first book of the Harry Potter series, with each set comprising one question and one correct answer.

Since SOTA DG method (Wang et al., 2023) utilized the MCQ dataset (Ren and Zhu, 2021) for training, it is treated as the in-domain dataset. We chose the MCQ test set, comprising a total of 259 multiple-choice question sets, as our testing dataset.

For the adaptability experiments, we used the MCQ dataset (Ren and Zhu, 2021) as the in-domain dataset, consisting of 259 multiple-choice question sets. The MEDMCQA dataset (Pal et al., 2022) served as the out-of-domain dataset, with 100 question sets selected to test the SOTA DG model's adaptability in less related contexts.

In the controllability experiment, we generated 100 questions using ChatGPT based on the first Harry Potter book, with the scope restricted to this book only.

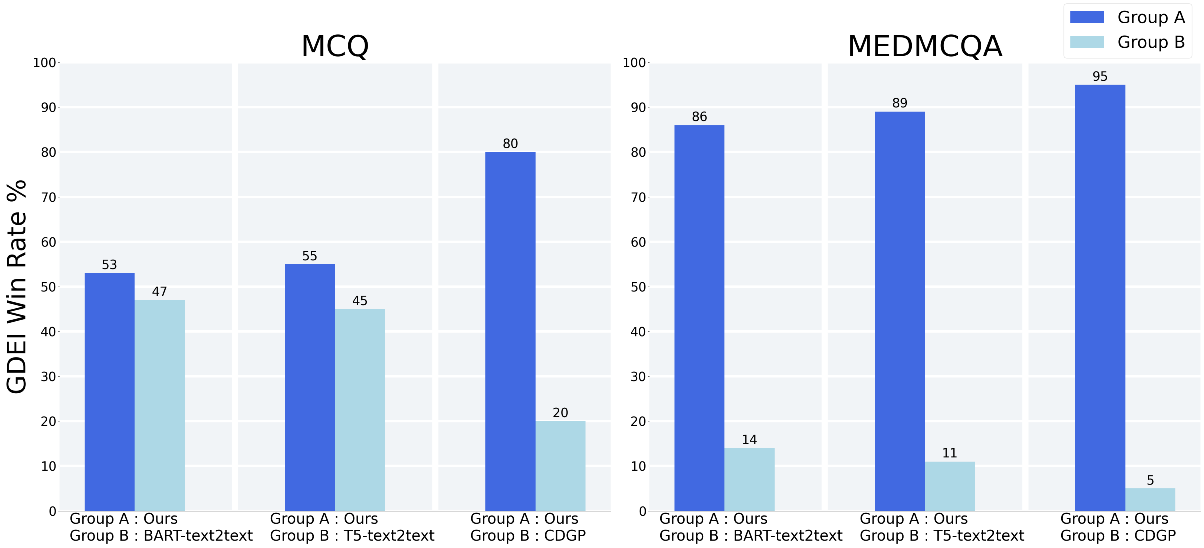


Figure 5: Experiment results of our method compared to DG-text2text(Wang et al., 2023) on both in-domain and out-domain datasets. While the margin of superiority in the in-domain dataset is modest, our method significantly outperforms the DG-text2text(Wang et al., 2023) in the out-domain dataset. This demonstrates the adaptability of our method across different domains.

5.4 Comparative Results Discussion

Our experimental analysis presents a comparative evaluation of our distractor generation method against SOTA DG model variants, namely T5 and BART, using two datasets: MCQ and MEDMCQA. The GDEI metric serves as our evaluation criterion, quantifying the effectiveness of the generated distractors.

5.4.1 Experiment : Adaptability

The overall result is shown in Figure 5.

- **Performance on MCQ Dataset:** In the MCQ experiment, our comparison targets two state-of-the-art training methods with different base models, namely T5-text2text and BART-text2text, as well as another method using the Pre-trained Language Model - CDGP (Chiang et al., 2022). In comparison with the state-of-the-art methods, our method achieved a GDEI score of 55, while the SOTA DG T5-text2text scored only 45. Our method also obtained a GDEI score of 53, whereas the SOTA DG BART-text2text scored only 47. In comparison with CDGP, our method achieved a score of 80, significantly surpassing CDGP's score of 20. This indicates a clear preference for the distractors generated by our method, suggesting that our approach produces more contextually relevant and challenging distractors.

- **Performance on MEDMCQA Dataset:** The MEDMCQA dataset further validates the superiority of our method. It attained remarkably high GDEI scores of 89 and 86, whereas the SOTA DG models scored significantly lower, with 11 for DG-T5 and 14 for DG-BART. In comparison with CDGP, our method achieved a score of 95, significantly surpassing CDGP's score of 5. The absolute contrast in the scores on this dataset emphasizes the robustness of our method in a specialized domain.

In practical educational settings, teachers across various fields need models to generate distractors. Insufficient model adaptability necessitates fine-tuning with specific distractor datasets for each field, which can be costly.

Combining results from both in-domain and out-domain datasets, our method slightly outperformed the SOTA DG model in GDEI scores within in-domain datasets but significantly surpassed it in the out-domain MEDMCQA dataset. This illustrates the limitations of traditional text-to-text methods on out-domain data, which lack adaptability. In contrast, our non-parametric architecture performs consistently across different domains without requiring fine-tuning. This stability and adaptability highlight our framework's alignment with practical educational applications, as it achieves strong performance across various domains using relevant reference articles, eliminating the need for additional manually annotated distractor datasets.

Generation Method	Distractor-wise	Question-wise
Llama 2 prompting	73%	58%
Llama 2 RAG prompting	79%	61%
Ours	100%	100%

Table 2: This table shows the probability that the generated distractors are within the specified range. The "All distractors" column indicates the probability that all 300 generated distractors are within the specified range. The "All set" column indicates the probability that all distractors for each of the 100 question sets are within the specified scope.

5.4.2 Experiment : Controllability

Performance on Harry Potter Novel dataset: Since the state-of-the-art (SOTA) distractor generation (DG) model was not trained on Harry Potter data, it struggles to generate relevant distractors for such content. To address this, we compared our extractive generation strategy with two prompting methods using Llama 2(Touvron et al., 2023) . The first method, Llama 2 prompting, involved directly inputting the question and answer to generate distractors. The second, Llama 2 RAG prompting, included a passage retrieved by BM25 with the correct answer.

We aimed to simulate realistic scenarios where teachers specify question scopes and tested our method with 100 questions from the first Harry Potter book, comparing it to the two Llama 2 strategies. Besides evaluating GDEI scores, we assessed how well the distractors matched the predefined scenario.

The results, presented in 6 and Table 2, show that although our method's GDEI scores are lower or only marginally higher compared to the state-of-the-art generation strategies, this is likely due to our method's constrained scope, which limits the generation of higher quality distractors outside this range.

Our method consistently met the specified scope conditions 100% of the time, in contrast to the Llama 2 prompting methods. With Llama 2 prompting, only 73% of distractors met the criteria of being from the first Harry Potter book, and Llama 2 RAG prompting improved this to 79%. When considering question sets, both prompting methods produced compliant distractor sets only 58% and 61% of the time, respectively.

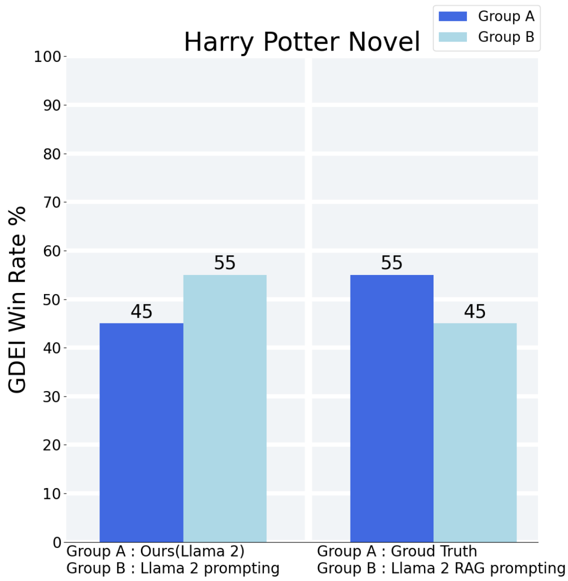


Figure 6: The comparative results of our method against two generation strategies of SOTA DG methods using llama 2 as the base model in Harry Potter Novel dataset.

These findings suggest that while large language models have parametric knowledge, they lack effective controllability for real educational scenarios. Our framework, despite lower GDEI scores, more effectively aligns with practical distractor design needs.

5.5 Case Study

5.5.1 MedMCQA case study:

In Table 3, we present a case study from the MedMCQA dataset. We can observe that in the first question, distractor set generated by our method ('perphenazine', 'penfluridol', 'chlorpromazine') is better for questions on refractory schizophrenia treatment as it includes drugs actually used for schizophrenia, requiring deeper knowledge to identify the correct answer. Distractor set generated by SOTA DG model('ibuprofen', 'tricyclics', 'Valium') includes less relevant drugs, making them less effective distractors.

In the second question, distractor set generated by our method is superior as it includes terms closely linked to protein folding, like 'Calnexin' and 'aspaate.' These are plausible distractors given their role in protein quality control. While 'Ribosome' is indirectly related, ('Xerophyte', 'Protist', 'Emb') lack relevance to protein folding, making distractor set generated by SOTA DG model less effective.

In the third question, distractor set generated by our approach not only exhibits a higher level of professionalism but also highlights the inadequacy of the SOTA distractor generation method in terms of knowledge in the out-domain scope. When encountering unfamiliar questions, the SOTA DG method may also produce redundant distractors.

5.5.2 Harry Potter Novel case study

In Table 4, we present a case study from the Harry Potter Novel experiment. In the first question example, distractor generated by our framework, is more effective because it includes items closely related to the Harry Potter universe that could be mistaken for something granting invisibility. The 'forgetfulness potion' and 'dragon egg' are magical but do not provide invisibility, making them plausible but incorrect. The 'mirror' is slightly less relevant but still within the magical realm. On the other hand, distractor generated using Llama 2 prompting, includes 'wand' and 'glasses,' which are relevant but less likely to be confused with an invisibility item, and 'crystal,' which is vague and less misleading. Distractor generated by our framework aligns better with real teacher practices as it requires students to distinguish between different magical items in the context of the story.

In the second question example, we observed that our framework produces distractors similar with those generated using the parametric knowledge of Llama 2. This demonstrates that our method, which relies on pseudo-answers extracted from the provided passages, yields results similar to those achieved with Llama 2's extensive parametric knowledge. These findings validate the effectiveness of our approach in using hard negative passages to generate distractors.

6 Conclusion

In conclusion, our DG method stands out for its unique features, notably the absence of fine-tuning requirements and the unprecedented control over question scope. Unlike other DG models, our approach operates efficiently without the need for extensive fine-tuning, allowing for swift adaptation across diverse domains by simply adjusting the knowledge corpus. The ability to selectively focus on generating questions within specific topic boundaries addresses the practical needs of educational contexts, providing educators with a level of control over question scope that conventional Text2Text models struggle to achieve. These distinctive characteristics position our DG method as a versatile and efficient tool for generating targeted and relevant questions across various subject areas.

7 Limitations

- **Assessment of Distractor Quality** Although the GDEI provides a more comprehensive evaluation of distractors, it may not capture all aspects of distractor quality, such as the potential for a distractor to reinforce common misconceptions or to be pedagogically useful.

- **Dependence on Quality of Corpus:**The quality and diversity of the generated distractors are directly tied to the richness of the knowledge corpus used. If the corpus is outdated, biased, or lacks depth, the distractors may not be as effective or may inadvertently introduce inaccuracies.

8 Appendix

A Ablation Study

A.1 Generation parameters

In figure 7, we tested three different k and h settings: k=60, h=5; k=100, h=3; and k=300, h=1. We observed that although the GDEI scores for the three settings were close when compared with SOTA-BART, the setting of k=300, h=1 significantly outperformed the other two settings when compared with SOTA-T5. We speculate that this is because we divided the corpus into small passages, each containing a limited amount of information. If h increases, it may result in meaningless distractor candidates. The optimal setting is to examine as many passages as possible and generate one distractor at a time for each passage.

A.2 Corpus chunk size

In the experiment in figure 8, we divided the corpus into chunks of varying sizes based on the number of sentences to examine the impact of chunk size on the generation results. We used the llama2-based Contextual Distractor Synthesizer with a setting of k=300 and h=1. Our data indicates that as the number of sentences per chunk decreases, the quality of the generated output improves. This outcome might be due to the extensive knowledge contained in each Wikipedia article, where each sentence potentially includes valuable words that can serve as distractor candidates. If the chunk size is too large, many potential distractors may be overlooked during generation. Conversely, smaller chunk sizes allow for a more detailed generation of high-quality distractors from each sentence within an article.

A.3 Ablation Study on different base model of the Contextual Distractor Synthesizer

We tested the impact of different base models for the Contextual Distractor Synthesizer on the generation results in figure 9. We fine-tuned Llama2-7b and Mistral-7b (Jiang et al., 2023) models for Contextual Distractor Synthesizer. Llama2-7b slightly outperformed Mistral-7b in both MCQ and MEDMCQA. We attribute this difference to the variations in the pre-training data of the lauguage model, which lead to differences in the models' s parametric knowledge. However, according to our proposed framework, as long as a Reading Comprehension model is well fine-tuned, even a smaller model can produce comparable generation results. With an adequately searchable corpus, high-quality distractors can be generated with our framework.

Algorithm 1 Distractor Generation Algorithm

Input: Knowledge corpus $\mathcal{C}$ consisting of text chunks, question stem Q, answer A to Q
Assume:
    (1) a document retriever $\mathcal{F}()$ for fetching text chunks from $\mathcal{C}$
    (2) an extractive document reader $\mathcal{R}()$ for extracting an answer from a given text and a question
procedure GENERATEDISTRACTORS($\mathcal{C}, Q, A$)
    $\kappa \leftarrow \mathcal{F}(Q)$ ▷ Fetching Top-k text chunks
    for all $C_i \in \kappa$ **do**
        if A is in C_i **then**
            Remove C_i from κ
        end if
    end for
    Let PAS be Pseudo Answer Set = $\{\}$
    for $C_i \in \kappa$ **do**
        $PseudoAnswer \leftarrow \mathcal{R}(C_i, Q)$
        Compute S_{retr} for $PseudoAnswer$
        Evaluate $PseudoAnswer$ with character-level ROUGE-L score
        if $PseudoAnswer$ has low Rouge similarity with all items in PAS **then**
            $PAS.add(PseudoAnswer)$
        end if
    end for
    EVALUATEDISTRACTORS(PAS, Q, A)
    return Top-n scored items in PAS as final distractor set for Q and A
end procedure
function EVALUATEDISTRACTORS(PAS, Q, A)
    for all $PA_i \in PAS$ **do**
        Compute $S_{confidence}$ for PA_i
        $score(PA_i) \leftarrow S_{retr} \cdot S_{confidence}$ ▷ Computing final score
    end for
    Sort PAS based on $score(PA_i)$ in descending order
    return Sorted PAS
end function

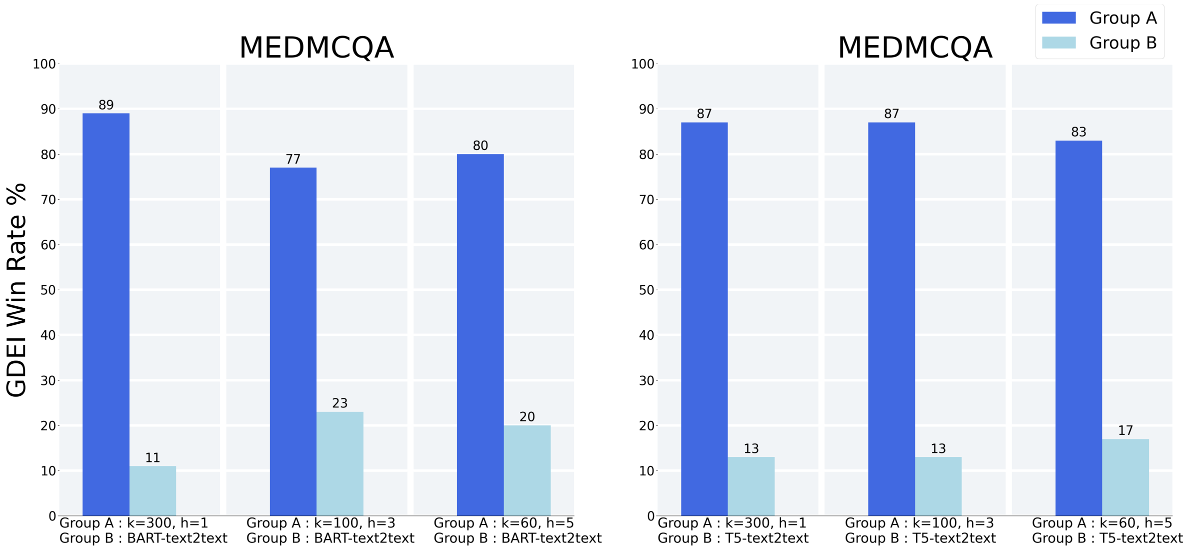


Figure 7: Experiment results of different settings of k,h parameters in comparison to DG-text2text (Wang et al., 2023) in MEDMCQA dataset.

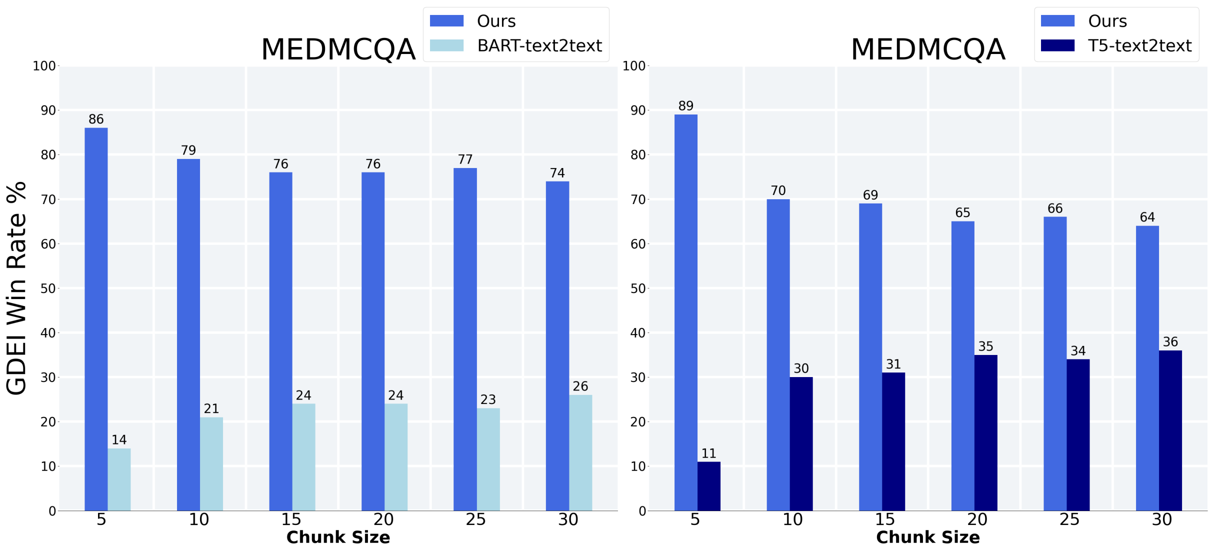


Figure 8: Different chunk sizes on the generation results in comparison to DG-text2text (Wang et al., 2023) in MEDMCQA dataset.

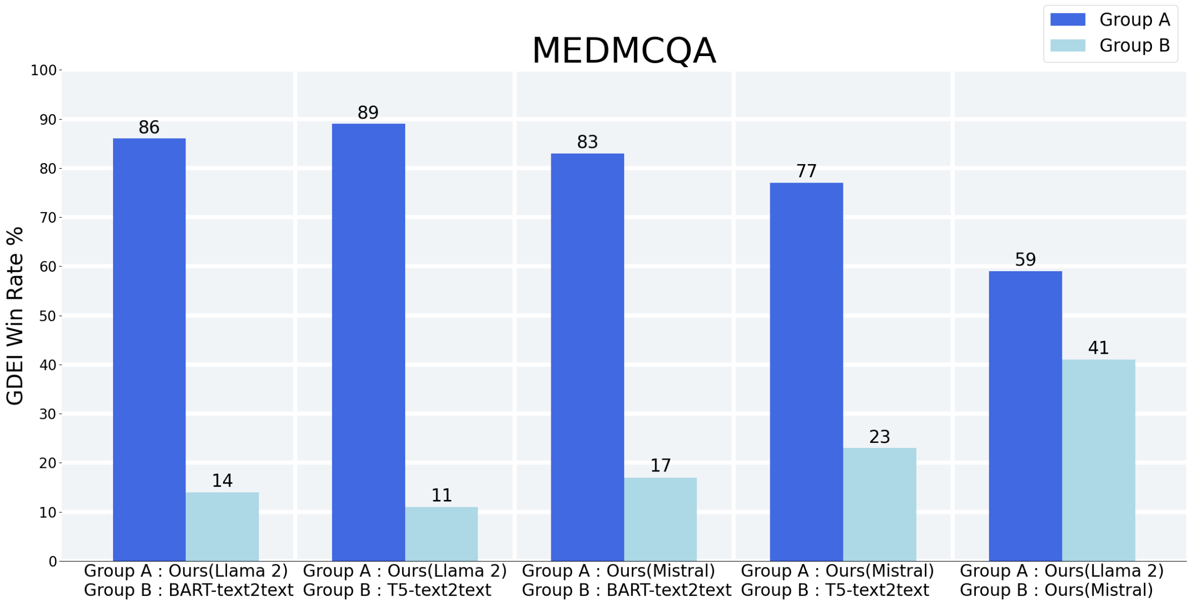


Figure 9: The performance of different base model on the Contextual Distractor Synthesizer in comparison to the SOTA DG model in MEDMCQA dataset.

Distractors are designed to be plausible but incorrect choices. Evaluate the quality of a distractor for the given sentence using the specified criteria.
Apply the following criteria for evaluation:
1. Exclusion from Correctness: distractors should not correctly fill the sentence.
2. Avoidance of Repetition: distractors should not be already mentioned in the sentence.
3. Relevance: distractors should be relevant to the question, not entirely unrelated or absurd choices.
4. Misleading: distractors should contain elements that may lead the test taker to select them incorrectly.
5. Alignment with Real Teacher Practices: Distractors must reflect the approach of real teachers in
crafting questions tailored for students.
Question: {question}
Answer: {answer}
Distractor set 1: { distractor set 1 }
Distractor set 2: { distractor set 2}
From a teacher's perspective, which of the following sets do you think is better: Distractor set 1 or 2.
Please output according to the following format.
reason: _(up to 100 words)
choose:set 1 or set 2 or both are same

Figure 10: Prompt for GPT-4 distractor evaluation for adaptability experiment.

You are the author of Harry Potter
Distractors are designed to be plausible but incorrect choices. Evaluate the quality of a distractor for the given sentence using the specified criteria.
Apply the following criteria for evaluation:
1. Exclusion from Correctness: distractors should not correctly fill the sentence.
2. Avoidance of Repetition: distractors should not be already mentioned in the sentence.
3. Relevance: distractors should be relevant to the question, not entirely unrelated or absurd choices.
4. Misleading: distractors should contain elements that may lead the test taker to select them incorrectly.
5. Alignment with Real Teacher Practices: Distractors must reflect the approach of real teachers in crafting questions tailored for students.
Question: {question}
Answer: {answer}
Distractor set 1: { distractor set 1 }
Distractor set 2: { distractor set 2}
From a teacher's perspective, which of the following sets do you think is better: Distractor set 1 or 2.
Please output according to the following format.
reason: _(up to 100 words)
choose:set 1 or set 2 or both are same

Figure 11: Prompt for GPT4 distractor evaluation for Controllability experiment.

References

Jun Araki, Dheeraj Rajagopal, Sreecharan Sankaranarayanan, Susan Holm, Yukari Yamakawa, and Teruko Mitamura. 2016. Generating questions and multiple-choice answers using semantic analysis of texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1125–1136.

Shang-Hsuan Chiang, Ssu-Cheng Wang, and Yao-Chung Fan. 2022. Cdgp: Automatic cloze distractor generation based on pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2020. A bert-based distractor generation scheme with multi-tasking and negative answer training strategies. *arXiv preprint arXiv:2010.05384*.

Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R Lyu. 2019. Generating distractors for reading comprehension questions from real examinations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6423–6430.

Qi Guo, Chinmay Kulkarni, Aniket Kittur, Jeffrey P Bigham, and Emma Brunskill. 2016. Questimator: Generating knowledge assessments for arbitrary topics. In *IJCAI-16: Proceedings of the AAAI Twenty-Fifth International Joint Conference on Artificial Intelligence*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Girish Kumar, Rafael E Banchs, and Luis Fernando D'Haro. 2015. Revup: Automatic gap-fill question generation from educational texts. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 154–161.

Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C Lee Giles. 2018. Distractor generation for multiple choice questions using learning to rank. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 284–290.

Chen Liang, Xiao Yang, Drew Wham, Bart Pursel, Rebecca Passonneaur, and C Lee Giles. 2017. Distractor generation with generative adversarial nets for automatically creating fill-in-the-blank questions. In *Proceedings of the Knowledge Capture Conference*, pages 1–4.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Hsien-Yung Peng, Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2022. Misleading inference generation via proximal policy optimization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 497–509. Springer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Siyu Ren and Kenny Q. Zhu. 2021. Knowledge-driven distractor generation for cloze-style multiple choice questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4339–4347.

Siyu Ren and Kenny Q Zhu. 2021. Knowledge-driven distractor generation for cloze-style multiple choice questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4339–4347.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Hui-Juan Wang, Kai-Yu Hsieh, Han-Cheng Yu, Jui-Ching Tsou, Yu An Shih, Chen-Hua Huang, and Yao-Chung Fan. 2023. Distractor generation based on text2text language models with pseudo kullback-leibler divergence regulation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12477–12491.

question	answer	Distractor
FDA approved drug for refractory schizophrenia **blank**?	Clozapine	ours: 1.perphenazine 2.penfluridol 3.chlorpormazine
		DG-text2text (Wang et al., 2023): 1.Ibuprofen 2.Tricyclic 3.Valium
Reverse folding of proteins is carried out by **blank**?	Chaperone	ours: 1.calnexin 2.ribosome 3.aspaate
		DG-text2text (Wang et al., 2023): 1.Xerophyte 2.Protest 3.Emb
Drug of choice of benzodiazepine poisoning is **blank**?	Flumazenil	ours: 1.midazolam 2.imidazenil 3.naloxone
		DG-text2text (Wang et al., 2023): 1.Alcohol 2.Cigarettes 3.Alcohol

Table 3: Case study of our work comparing to the SOTA DG model in MEDMCQA dataset

question	answer	distractor
What magical item did Harry receive that made him invisible?	Invisible cloak	ours: 1. forgetfulness potion 2. dragon egg 3. mirror
		Llama 2 prompting: 1. crescent 2. crystal 3.wand
Who is the caretaker of Hogwarts?	Filch	ours: 1. rubeus hagrid 2. dumbledore 3. professor snape
		Llama 2 prompting: 1. hagrid 2. dumbledore 3. mcgonagall

Table 4: Case study of our work comparing to the SOTA DG model in Harry Potter Novel dataset

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2017. Large-scale cloze test dataset created by teachers. *arXiv preprint arXiv:1711.03225*.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14506–14514.

Xiaorui Zhou, Senlin Luo, and Yunfang Wu. 2020. Co-attention hierarchical network: Generating coherent long distractors for reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9725–9732.

Combining Topic and GNN Models for Text Classification

Lawrence Y. H. Low **Yen-Tsang Wu** **Jenq-Haur Wang**
Department of Computer Science and Information Engineering
National Taipei University of Technology
Web Information Retrieval Lab
t113999402@ntut.org.tw buddyswu@gmail.com jhwang@ntut.edu.tw

Abstract

Deep Learning (DL) models in Natural Language Processing (NLP) are divided into two categories: Sequential-based and Graph-based. More recently, Sequential models use Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Bidirectional Encoder Representations from Transformers (BERT). Researchers have also applied graph-based model to NLP, building graphs to learn the semantic features of text. In this study, we propose an efficient graph-based model called TopicGraph by combining Topic modeling and Graph Neural Network (GNN) which allows the reusing of the trained model to easily build graphs for inference since words from a topic has been learned during the training phrase. We use four different English datasets for our experimentation. Comparing with other sequential and graph-based models, the accuracy of our proposed model TopicGraph on MR, R8, R52, and Ohsumed datasets achieve 79%, 97%, 94%, and 72%, respectively. By adding a Subgraph-wise sampling technique, GraphSAINT for feature extraction, together with BERT as a multi-class classifier, we further improve the accuracy of our TopicGraph model of MR by 12%, R8 by 0.08%, R52 by 1.71% respectively. This clearly demonstrated the effectiveness of our proposed approach.

Keywords: Topic model, Graph neural network, Subgraph-wise sampling, Text classification

1 Introduction

With the advancement of Large Language Models (LLMs) and Generative Artificial Intelligence (AI), more works have been researched using these techniques in the field of Deep Learning (DL) recently, especially in the area of Natural Language Processing (NLP). However, it is also equally important to research into the primary task of text classification to understand the semantic meaning of text in order to accurately and efficiently classifying them.

The traditional text classification method usually is using bag-of-words to extract the semantic and linguistic meaning of text by using term frequency (TF). Topic Modeling is a traditional Machine Learning (ML) method which try to classify text document based on topics that are extracted from a text. The above two methods do not take into consideration of word position within a document, but focus instead on the word occurrence.

More advance Neural Network (NN) techniques was employed to look into this limitation by applying CNN (Kim, 2014), RNN (Liu, Qiu, & Huang, 2016) and BERT (Devlin et al., 2019) framework with the advancement of DL technology. RNN and BERT both considered when and where a text appears in a document in order to extract more vital features from a text, this can further improve the text classification accuracy.

CNN was initially used for image classification by using the convolutional layer to extract the image features. In the field of NLP, CNN can be applied to extract the local information of N-gram features in a text document. More recently, Graph Convolutional Neural Network (GCN) (Kipf and Welling, 2017) is a GNN technique combining both CNN and NN in order to capture the global information or Co-occurrence features in a text document using graph-based method.

In this study, we propose a graph-based model called TopicGraph by combining both Topic model and Graph Neural Network (GNN) model to classify text.

2 Related Work

2.1 Dependency parser graph-based model

Chou et al. (2020) introduced a dependency parser graph-based model for text classification. The researchers used RNN to calculate the text features in a text sequence. All texts are treated as a graph node by using the node modified relationship between the texts to build a graph. Their Dep-GAT-root model, combined with Graph Attention Network (GAT) and a dependency parser for graph construction, extracted text features to perform English text classification with good results.

2.2 Gated Recurrent Unit (GRU)

The Gated recurrent unit (GRU) is a form of RNN (Cho et al., 2014). This hidden unit consists of 2 gates, a reset gate and an update gate. These 2 gates control the information flowing within the hidden cell, either to update the hidden state using the update gate or forgetting the previous state using the reset gate. The update gate will determine if a hidden state requires an update with a new hidden state, whereas the reset gate determines the exemption of a previous hidden state. This GRU unit is slightly different from Long Short-Term Memory (LSTM) which has 4 gates and a memory cell (Graves, 2012). Since GRU has only 2 gates, it has fewer parameters than LSTM, hence it is more efficient and requires less computational powers.

2.3 Latent Dirichlet Allocation (LDA) for Topic Modeling

The Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) is a generative probabilistic topic model that can discover hidden latent topics from a corpus that is categorized by their word distribution. In the LDA model, the nodes represent random variables whereas the edges represent the dependent relationship in between the nodes and the edges. In general, the generative process of LDA is defined as follows:

1) For each topic k, get a distribution over the vocabulary V.
2) For every document d, get a distribution over topics (per-document topic proportion).
3) With every document d, for a word w within document d, get a topic assignment (per-word topic assignment).
4) With every document d, for a word w within document d, get a word from the vocabulary V.

2.4 Graph Convolutional Network (GCN)

A graph is a data structure type that is made up of 2 components, a node and an edge. The most basic graph can be represented as $G = (V, E)$ where V is the set of nodes, and E is the edges or links between them. Edges between the nodes are identified as directed if there are any directional dependencies in between the nodes. Edges are identified as undirected when there are no directional dependencies.

Graph Neural Network (GNN) is divided into four main categories (Wu et al., 2019). They are Recurrent Graph Neural Networks (GNN), Convolutional Graph Neural Networks (CGNN), Graph Autoencoders, and Spatial-Temporal Graph Neural Networks. In this paper, we focus on works in the category of Graph Convolutional Network (GCN), Kipf and Welling (2017) introduced a scalable graph architecture that is used for semi-supervised learning on graph-structured data. This architecture is based on the CNNs which operates directly on graphs. The GCN model can scale linearly based on the number of graph edges. More importantly, the hidden layer representations that used to encode both the local graph structure and features of nodes are learnable.

For GCN to be scalable and computationally efficient, GCN uses layer-wise propagation rule which is based on the first-order approximation of spectral convolutions on graphs. GCN is a transductive model that is trained on a specific graph and makes predictions for the nodes within that graph. The model has access to the entire graph structure during both training and inference. The key characteristic is that the model can leverage information from the entire graph, including all nodes and edges, to learn embeddings and make predictions. However, this characteristic has its limitation; this GCN model cannot be applied to new data that was not seen during training.

2.5 Graph Attention Networks (GAT)

To overcome the transductive nature of the GCN model, adding an attention layer to the GCN model proved to be effective. Veličković et al. (2018) presented the Graph Attention Networks (GAT) model which is a combination of a GNN with an attention layer architecture.

GAT leveraged on the masked self-attentional layers in the architecture to address the drawbacks based on graph convolutions. The implementation of the attention mechanism enables different weights to be assigned to the different nodes in a neighborhood. GAT model is parametrized by a weight vector with LeakyReLU activation and the GAT multiheaded attention have nodes with different weights assignment on its neighborhood. The key difference between GAT and GCN is how the information from one-hop neighborhood is combined.

2.6 Word2Vec

Mikolov et al. (2013) proposed the 2 new model architectures known collectively as Word2Vec to create dense representation of word embeddings. Word2Vec is a continuous vector representation of words for any set of corpora. However, it does not consider the positions or relationships between the words and 1 word is usually a representation in Word2Vec.

Continuous Bag-of-Words (CBOW) is an architecture in Word2Vec. CBOW uses both words in the past and words in the future for word embeddings. Skip-gram is the other architecture in Word2Vec. Instead of using past and future words for word embeddings in the case of CBOW, Skip-gram focuses on a centralized word to predict the surrounding words. In the CBOW model, the surrounding context words representations are used to predict a target word. For example, given a series of context words “the brown bear up the tree”, CBOW model will use these context words to predict the target word such as “climbs”, “hops” … based on the model parameters.

Different from CBOW, Skip-gram model used the target context word to predict its surrounding context words. For example, given the context word “climbs”, Skip-gram model will use the target context word to predict its surrounding words such as “the”, “brown”, “bear”, “up”, “the”, “tree” … based on the model parameters.

2.7 Subgraph-wise Sampling: GraphSAINT

Graph sampling based inductive learning method (GraphSAINT) introduced by Zeng et al. (2020) is a Subgraph-wise sampling technique. The purpose of GraphSAINT is to provide a scalable algorithm for large scale GNN models training by utilizing small subgraphs sampling.

First, GraphSAINT builds minibatches for training by sampling the training graph instead of using the nodes or edges across the full GCN layers. A complete GCN is constructed from the sampled subgraph for each iteration. In this way, GraphSAINT ensures that a fixed number of well-connected nodes are connected throughout all the layers.

The training process can be defined in 3 steps as follows:

1) Select subgraphs randomly.
2) Construct and run GNN only on the subgraph level.
3) Repeat recursively.

In this paper, we generated the subgraphs using 3 random sampling techniques in GraphSAINT (Zeng et al., 2020), which are Node sampler, Edge sampler and Random Walk sampler (Wang et al., 2019). The random walk sampler begins at a randomly selected starting node, which is selected from the set of root nodes.

The process of random walking steps can be illustrated as follows:

1) Step 1: Randomly select a starting node as the starting point, move to a neighboring node.
2) Step 2: After reaching the neighboring node in Step 1, move to another neighboring node.

We select a shorter random walk as it will capture nearby local neighborhood information around the starting nodes. This can be useful for capturing fine-grained, local structures in the graph. GraphSAINT increases the overall performance of a graph model both in terms of accuracy and training time.

2.8 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a Language Model (LM) created by Google researchers, (Devin et al., 2018) to improve the fine-tuning approaches used in pre-trained language representations, such as OpenAI Generative Pre-trained Transformer (GPT). OpenAI GPT uses a unidirectional architecture, where only the previous tokens are the main focus in the self-attention layers of the Transformer (Vaswani et al., 2017).

To mitigate this unidirectionality problem, BERT introduces a “Masked Language Model” (MLM) pre-training method that masked the input tokens randomly. The goal is to predict the original Vocabulary ID of the masked word at its context level only, which is different from standard LM architecture using a left-to-right unidirectional LM pre-training approach.

BERT uses same architecture in both pre-training and fine-tuning approaches in pre-training the LMs. All the different NLP down-stream tasks utilize the same pre-trained model parameters to initialize the model.

During fine-tuning process, all fine-tuning parameters are being fine-tuned. A special token [CLS] is added in front of each text input, and another special separator token [SEP] is used for separating 2 texts, such as in the questions-answers pairs.

To enable BERT to handle various NLP tasks, a BERT input or output representation can represent both a sentence and a pair of 2 sentences in 1 single token sequence clearly. A sentence in a BERT model refers to an arbitrary span of contiguous text, instead of an actual linguistic sentence. A typical BERT input representation is the summarization of its token embeddings, segmentation embeddings, and position embeddings.

BERT MLM method is used to pre-train a deep bidirectional Transformer, which itself is an Encoder only Transformer which is different from the standard Transformer that consists of a Decoder. BERT also introduces a "Next Sentence Prediction" (NSP) task that can jointly pre-trains text-pair representations. BERT improved in 11 state of the art (SOTA) NLP tasks, which include advancing the GLUE score to 80.5%, MultiNLI accuracy to 86.7%, SQuAD v1.1 question answering Test F1 to 93.2 and SQuAD v2.0 Test F1 to 83.1 (Devin et al., 2018).

3 Methodology

We divided our proposed methodology into 5 major processes as shown in Figure 1 below (Chou, 2020). The 5 processes are Topic modeling, Graph Construction, Feature Extraction, Feature Fusion, and Text Classification.

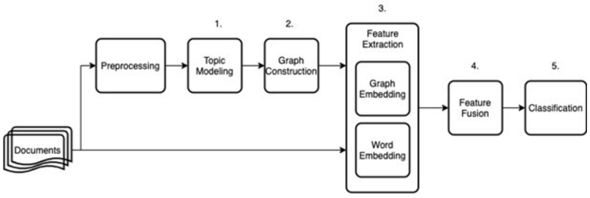


Figure 1. Methodology of TopicGraph model.

The data go through a round of data preprocessing before the 5 processes. We filtered out the lesser important words from the dataset by removing the punctuations and using stopwords from spaCy. After data preprocessing, we train our topic model using a LDA model to extract the topics distribution and construct a word-topic graph. For feature extraction, we utilize the GCN-GAT framework to extract the graph embedding whereas for word embedding, we use Word2Vec and RNN. By concatenating both the word-topic graph and word embedding during feature fusion, we extract the final embedding to perform supervised text classification.

3.1 Topic Modeling

For Topic Modeling, we train a LDA model to extract the most relevant topics from a text document. For example, the word "Genetics" may be the most significant word in a text document containing the various text consisting of "human", "genome", and "DNA" and so forth.

The LDA model is able to find the topic and its related relevant words. These features are a kind of global information of a text document which we can treat it as a form of word co-occurrence. For example, when we have words like "human", "genome", and "DNA" appearing in a text document at the same time, we may assume that the text document is related to the topic "Genetics". In this study, we try to introduce these features for constructing a graph.

3.2 Graph Construction

To construct our graph, first we input our text sequence into the LDA model to find the most probable topic for each word, *p(topic|word)* in the text sequence as shown in the Figure 2 below (Chou, 2020).

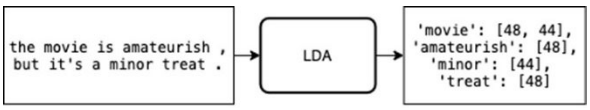


Figure 2. LDA topic modeling example.

For our example in the above Figure 2 which we set a maximum K topic of 50, the input text sequence "the movie is amateurish, but it's a minor treat." may generate for the word "movie" with 2 topics [48, 44], "amateurish" with 1 topic [48] and so forth.

Based on the above results, we construct a graph using the topics and words as nodes, and the connected relationships between the topics and words as edges as shown in Figure 3 below (Chou, 2020).

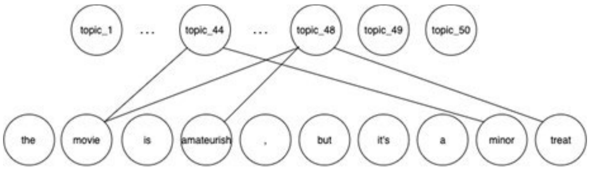


Figure 3. LDA topic modeling nodes example.

From the above example in Figure 3, we can see that the relationship between the words in a text document. At the same time, we can further observe that for the word node, its 1-hop neighborhood is its relationship with its topic feature whereas the 2-hop neighborhood is its relationship with other related words with the same topics in the text sequence. We hope that by using these 2 connected relationship features, the model can learn new features from the text.

We use our Topic model that computes topic classification and utilizes the topic node which has an edge with the word node. A text sequence is converted into word embeddings using a pre-trained Word2Vec file before inputting into the bi-directional RNN layer with GRU activation unit. Following that, the topics-word model is computed using LDA with to generate the topics-word embeddings.

At the same time, the graph embeddings will be generated by a 2-layer GCN together with a GAT layer. Once this is done, both the word and graph embeddings go through a fully connected linear layer before concatenating to generate the final output, which is a 100-dimensional word embedding file.

The text input sequence goes through several transformations in the TopicGraph model, which is listed as follows:

1) Extract forward and backward relationship of the text sequence using GRU.
2) Construct node embeddings using topic features from LDA and sequence features from GRU.
3) Reduce the GRU concatenated hidden features dimensionality using linear layers.
4) Extract the graph node embeddings from GRU concatenated hidden features using GAT layers and graph linear layer.
5) Concatenate the sequence and graph embeddings without the topic nodes.
6) Get the unnormalized output logits after the merged embeddings passed through a linear layer.
7) The output logits are aggregated to produce the final prediction output.
8) Compute the topic-specific output logits.
9) The model output is the final prediction output logits and the topic-specific logits.

3.3 Feature Extraction

Our feature extraction process consists of 3 methods as follows:

1) Using a pre-trained Word2Vec model directly.
2) Using RNN to find word embeddings that include contextual relationships.
3) Using GNN-GAT framework to find word embeddings.

For RNN feature extraction, we used a Bidirectional RNN (Schuster and Paliwal, 1997) to extract the contextual relationships of a word in text sequence as shown in Figure 4 below (Chou, 2020). In this study, each input text sequence is a vector converted by Word2Vec.

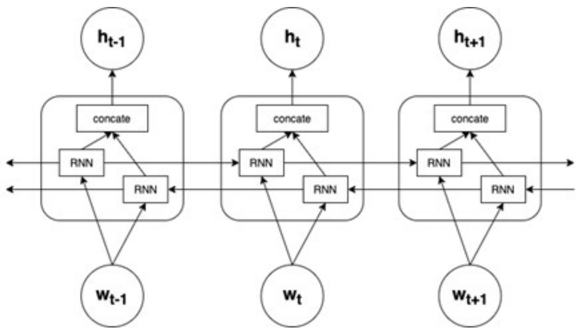


Figure 4. Schematic diagram of Bidirectional RNN.

GAT is a NN network that is applied in a graph architecture. It combines GCN with a self-attention mechanism (Bahdanau et al., 2015; Vaswani et al., 2017). When calculating the weights of the nodes in a graph, GAT focus on the importance of its neighboring nodes to update the weights of its own node, as shown in Eq. (1) and Eq. (2) below (Chou, 2020):

$$X' = GAT(X),\ x_i' = \alpha_{i,i}\Theta x_i + \sum_{j\in N(i)} \alpha_{i,j}\Theta x_j \quad (1)$$

$$\alpha_{i,j} = \frac{\exp\left(LeakyReLU(\vec{a}^T[\Theta x_i || \Theta x_j])\right)}{\sum_{k\in N_i}\exp\left(LeakyReLU(\vec{a}^T[\Theta x_i || \Theta x_k])\right)} \quad (2)$$

As shown in Figure 3 above, each text sequence is construct into a word-topic graph structure. The nodes are divided into two categories, topic nodes and word nodes. The features of the word nodes are used as the initial features by the vector calculated by Bidirectional RNN whereas the topic nodes are encoded with one-hot encoding. Since both topic nodes and word nodes are different in their features, different linear transformations applied here, as such Eq. (1) above is slightly modified into Eq. (3) as shown below (Chou, 2020):

$$x_i' = \alpha_{i,i}\Theta x_i + \sum_{j\in N(i)} \alpha_{i,j}\Theta x_j\,, \Theta = \begin{cases}\Theta_{word}, & x \text{ is word node}\\ \Theta_{topic}, & x \text{ is topic node}\end{cases} \quad (3)$$

Unlike RNN, which captures text features from the context, we expect GAT to capture different text features. From Figure 3 above, the text under the same topic will add topic features, such as "maturish" and "treat". As compared with the other words, "movie" belongs to two topics, so it will have a connection between the two topic features. Because the graph is undirected, the relationship between the topic feature and text features are also added. For example, topic_48 in Figure 3 above will add three text features. Although LDA can calculate the text distribution under each topic which is the importance of a word under a certain topic, we do not use the calculated values obtained by LDA, but instead we use GAT to learn new weights, which is more relevant in our target dataset.

3.4 Feature Fusion

The overall architecture of TopicGraph is as shown in Figure 5 below (Chou, 2020).

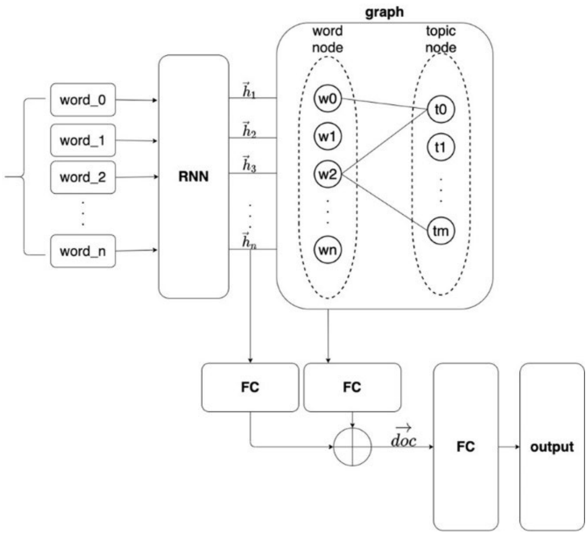


Figure 5. Overall Architecture of TopicGraph.

During feature fusion, we first extracted the vectors h of all words in a text sequence, n is the total number of words in the text sequence. We input the vectors in two directions into the model. First, the vectors are inputted into the graph that we constructed which allow each text to learn a new vector. Then we output the newly learned graph features into a Fully connected layer (FC). In this step, we define the node characteristics x in Eq. (4) and Eq. (5) as follows (Chou, 2020):

$$x_i = \vec{h}_i \quad \forall i \in \{1, 2, \dots, n\},\ x_j = \overrightarrow{topic_k} \quad \forall j \in \{n+1, n+2, \dots, n+k\} \quad (4)$$

$$g = tanh\,(\Theta_{graph}\, GAT(x)),\quad g = \{\vec{g}_1, \vec{g}_2, \dots, \vec{g}_{n+k}\} \quad (5)$$

GAT(x) is a function for updating node features, and the algorithm is as shown in Eq. (3) above.

The other direction of the inputted vector, the learned text feature calculated from RNN is directly connected to FC, as shown in Eq. (6) as follows (Chou, 2020):

$$h' = tanh(\Theta_{\text{rnn}}\, h) \quad (6)$$

According to Eq. (5) and Eq. (6), we connect a layer of FC to re-extract the features learned by RNN and GAT, then we obtain two different vectors h' and g. In order to obtain the vectors of the entire text sequence, we concatenate h' and g together and get the average value. In this study, only the newly extracted n word embeddings are used. Although using the average value cannot capture important features as well as the attention technique, it has good results in terms of performance and efficiency with no additional parameters required (Shen et al., 2018), as shown in Eq. (7) as follows:

$$\overrightarrow{doc} = \frac{1}{n}\sum_{i=1}^{n} concatenate\left(\overrightarrow{h'}_i, \vec{g}_i\right) \quad \forall i \in \{1,2,\ldots,n\} \quad (7)$$

3.5 Text Classification

The training parameters for the entire model are calculated using the Gradient Descent method. In this study, we use cross-entropy as our loss function for text classification. We need to minimize two losses, one is the Softmax loss and the other is the loss that are used for Topic classification. We hope that the trained topic features can capture some important semantic features that is useful in text classification.

4 Experiments

4.1 Dataset

We use the 4 datasets used by TextGCN model (Yao et al., 2019) which are MR, R8, R52 and Ohsumed for model performance and comparison purposes. All the datasets consist of English paragraphs with different classification labels that are meant for different NLP subtasks.

4.2 Model Evaluation

In order to evaluate the effect of the topic model on the entire model, we use two different evaluation methods. Topic coherence score is used in the topic model to evaluate the quality of the topic whereas for text classification, we use accuracy for classification evaluation. At the same time, in order to determine the stability of the model, t-test was used for statistical testing.

4.3 Experimental Settings

For topics modeling, we set the maximum number of topics to 100, with an interval of 10 plus the number of classes in the dataset. This will allow us to observe the effect of different number of topics on the dataset. The maximum epoch is set to 200 with early stopping level set at 10 epochs. The Learning Rate is set to 0.001 which is the result of preliminary observation during the experiment where we observed that validation loss decreases steadily. The number of graph convolution layer is set to 2 which we referenced from other studies (Hamilton et al., 2017; Kipf and Welling, 2017; Veličković et al., 2018).

4.4 Experimental Results

We show the model comparison of text classification results between our proposed model in this paper with other models (Chou et al., 2020). The comparison result is as shown in Table 1 below.

Model	**R8**	**R52**	**Ohsumed**	**MR**
TF-IDF+LR	0.9347	0.8695	0.5466	0.7459
LDA+LR	0.8149	0.7118	0.2537	0.5371
CNN-non-static	0.9571 ± 0.0052	0.8759 ± 0.0048	0.5844 ± 0.0106	0.7775 ± 0.0072
Bi-LSTM	0.9631 ± 0.0033	0.9054 ± 0.0091	0.4927 ± 0.0107	0.7768 ± 0.0086
PV-DBOW	0.8587 ± 0.0010	0.7829 ± 0.0011	0.4665 ± 0.0019	0.6109 ± 0.0010
PV-DM	0.5207 ± 0.0004	0.4492 ± 0.0005	0.2950 ± 0.0007	0.5947 ± 0.0038
PTE	0.9669 ± 0.0013	0.9071 ± 0.0014	0.5358 ± 0.0029	0.7023 ± 0.0036
fastText	0.9613 ± 0.0021	0.9281 ± 0.0009	0.5770 ± 0.0049	0.7514 ± 0.0020
SWEM	0.9532 ± 0.0026	0.9294 ± 0.0024	0.6312 ± 0.0055	0.7665 ± 0.0063
Bi-GRU-avg	0.9702 ± 0.0036	0.9235 ± 0.0072	0.6135 ± 0.018	0.7844 ± 0.0057
LEAM	0.9331 ± 0.0024	0.9184 ± 0.0023	0.5858 ± 0.0079	0.7695 ± 0.0045
Text GCN	0.9707 ± 0.0010	0.9356 ± 0.0018	0.6836 ± 0.0056	0.7674 ± 0.0020
Dep-GAT-root	0.9654 ± 0.0025	0.9263 ± 0.0062	0.6194 ± 0.0118	0.7942 ± 0.0059
Dep-GAT-avg	0.9611 ± 0.0075	0.9229 ± 0.0066	0.5630 ± 0.0146	0.7839 ± 0.0029
Ours				
TopicGraph	**0.9754 ± 0.0020**	**0.9427 ± 0.0048**	0.6362 ± 0.0104	**0.7933 ± 0.0041**
TopicGraph (without GRU)	0.9146 ± 0.0023	0.8022 ± 0.0096	**0.7271 ± 0.0121**	0.7573 ± 0.0091

Table 1: Classification model comparison.

The results show that our models can achieve good results on R8, R52 and MR datasets with an accuracy score of 0.9754, 0.9427 and 0.7933 respectively. RNN can get good results on shorter texts (Cho et al., 2014) and the features learned by the graph model help to improve text classification as well.

However, our TopicGraph model did not perform better than the TextGCN model in the Ohsumed dataset. The reason may be that our proposed model TopicGraph is not effective in classifying longer texts. In order to test whether RNN has indeed lower our model text classification performance, we did an ablation test to remove the GRU module from the overall model as shown in Figure 6 below (Chou, 2020).

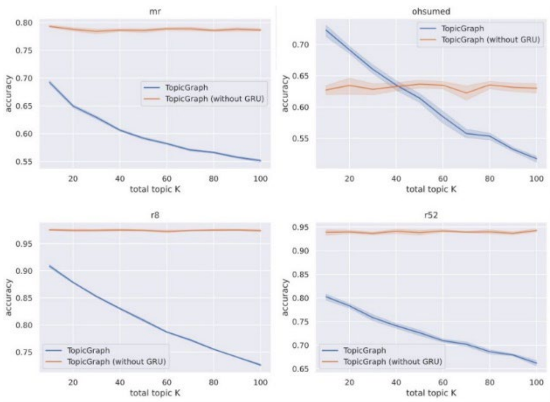


Figure 6. Ablation test by removing GRU.

From the above Figure 6, the TopicGraph (without GRU) model is added with sequential embedding of the RNN model. As compared with the TopicGraph model, we observed there is an increase in overall accuracy of the entire model and at the same time, the entire model accuracy is not affected by the numbers of topics in the Topic model.

Whereas for the Ohsumed dataset, we observed there is a decrease in the accuracy as compared to the TopicGraph model. However, if we removed the RNN embeddings, we observe that there is a significant improved results from the previous accuracy score of 0.6362 to 0.7271.

From these observations, we realized that our proposed model performed better on shorter text dataset R8, R52 and MR but not on longer text dataset Ohsumed. This also shown that RNN models can RNN perform better on shorter texts (Cho et al., 2014).

Although using our TopicGraph (without GRU) can improve in performance on Ohsumed dataset, but it could not get better classification results when comparing with TextGCN model which uses word co-occurrence and word-document relation to build graphs.

This may be due to the following:

1) Our TopicGraph (without GRU) model does not contain sequential relationship in between texts.
2) Our TopicGraph (without GRU) model does not obtain good topic words on shorter text document.

4.5 Further Improvement on TopicGraph Model

To improve our model performance, we use the subgraph-wise sampling technique, GraphSAINT (GS) to extract additional contextual features. We then use the newly extracted features to input into a fine-tuned BERT base model (uncased) with 2 NN layers for text classification with improved results. We show our modified TopicGraph-GS-BERT model results as shown in Table 2 below.

From the results in the Table 2 below, we can see there are improvement in performance for shorter text dataset R8, R52 and MR by 0.08%, 1.71% and 12% respectively. Although the modified TopicGraph-GS-BERT model increased in performance against TopicGraph model by 5.06%, but it performed lower than the TopicGraph (without GRU) by 1.01%. We think this is due to our TopicGraph model low ability to classify longer text. The (-) in Table 2, Table 3 below denote that the corresponding GS sampler was not used as the extracted features are not the optimal out of the 3 GS samplers, and (*) denote no results.

Model	**R8**	**R52**	**Ohsumed**	**MR**
TextGCN	0.9707 ± 0.0010	0.9356 ± 0.0018	0.6836 ± 0.0056	0.7674 ± 0.0020
Dep-GAT-root	0.9654 ± 0.0025	0.9263 ± 0.0062	0.6194 ± 0.0118	0.7942 ± 0.0059
Dep-GAT-avg	0.9611 ± 0.0075	0.9229 ± 0.0066	0.5630 ± 0.0146	0.7839 ± 0.0029
PMI-GAT-avg	0.9667 ± 0.0041	0.9256 ± 0.0096	0.5512 ± 0.0578	0.7606 ± 0.0059
Ours				
TopicGraph (without GRU)	0.9146 ± 0.0023	0.8022 ± 0.0096	**0.7271** ± 0.0121	0.7573 ± 0.0091
TopicGraph	0.9754 ± 0.0020	0.9427 ± 0.0048	0.6362 ± 0.0104	0.7933 ± 0.0041
TopicGraph-GS_{rw}-BERT	0.9690	0.9578	0.6664	**0.9133**
TopicGraph-GS_{node}-BERT	**0.9762**	**0.9661**	-	-
TopicGraph-GS_{edge}-BERT	-	-	0.7170	-

Table 2: Classification model comparison 2.

We conduct an ablation test to check the if using GraphSAINT to extract additional contextual features and using a fine-tuned BERT model for text classification did indeed impact the overall model performance. We show our ablation test results as shown in Table 3 below.

Model	R8	R52	Ohsumed	MR
Ours				
TopicGraph (without GRU)	91.46	80.22	**72.71**	75.73
TopicGraph	97.54	94.27	63.62	79.33
TopicGraph-GS_{node}	95.98	90.07	51.30	75.32
TopicGraph-GS_{edge}	*	*	51.62	*
TopicGraph-GS_{rw}	86.48	85.05	45.96	77.04
TopicGraph-BERT	96.52	94.50	65.76	86.04
TopicGraph-GS_{rw}-BERT	96.90	95.78	66.64	**91.33**
TopicGraph-GS_{node}-BERT	**97.62**	**96.44**	-	-
TopicGraph-GS_{edge}-BERT	-	-	71.70	-

Table 3: TopicGraph-GS-BERT Ablation test.

We made the following observations from the ablation test above:

1) The TopicGraph-GS (without TopicGraph and BERT) model did not perform better than the TopicGraph model. At the same time, the accuracy scores of TopicGraph-GS-BERT also dropped by 1.64% to 20.08% for the 4 datasets. This show that TopicGraph and BERT contain relevant and important semantic information.
2) The TopicGraph-BERT (without GS) model perform better in 3 of the datasets except R8 dataset. At the same time, the accuracy scores of TopicGraph-GS-BERT also dropped by 1.1% to 5.94% for the 4 datasets. This show that GS is an important component of the modified TopicGraph-BERT model.
3) The TopicGraph-GS-BERT model perform better in 3 datasets except Ohsumed dataset. At the same time, the accuracy scores of TopicGraph-GS-BERT increased by 0.08% to 12% for the 3 datasets.

From the ablation test results, TopicGraph-GS-BERT model did indeed improve the overall performance of the TopicGraph on shorter text dataset R8, R52 and MR, but not in the longer text dataset Ohsumed.

5 Conclusion

In this study, we combine topic model and GNN to classify text. We first use the topic model to capture the co-occurrence features of text. Then we represent both text and topics as nodes in a graph, and combining the features of text and topics to learn new word embeddings. Because our TopicGraph model considered the text as a whole unit when constructing a graph, we can easily add a sequential model to capture additional contextual features and rebuilding of the graph is not required when predicting new data.

We also observe the impact of topics on text classification performance. As the number of topics increases, the overall performance of text classification and topic coherence will also decrease. For GNN, increasing the number of convolution layers will degrade the model performance and using a 2-hop neighborhood feature usually has a better text classification result. Adding contextual features will achieve a better accuracy for shorter text datasets, such as MR, R8 and R52. However, for longer text dataset such as Ohsumed dataset, we can focus on the co-occurrence features and graph features alone.

In order to improve our TopicGraph model performance, we add the subgraph-wise sampling technique, GraphSAINT to extract new features. Using the new features with a fine-tuned BERT-NN model, we are able to further improve our model performance.

One particular weakness in the proposed model is the lack of model comparison with other SOTA models. We intend to look into various established transductive and inductive graph benchmarks for model comparison as one of our main focuses for our future works.

Another shortcoming is the lack of visual proofs and clear visibility of the proposed model strengths in our proposed model. We also noticed our proposed model weakness in classifying longer text document. The other focuses in our future works will be rectification to these shortcomings to further enhance the model capability.

Since our proposed model work on four datasets, testing our proposed model framework on different datasets will be another challenge which we will focus in the future.

Acknowledgments

The authors would like to thank the supports from the National Science and Technology Council, Taiwan under the grant numbers: NSTC113-2221-E-027-096, and NSTC113-2634-F-027-001-MBK. The authors acknowledge and thank Kuan-Hsun Chou for his contribution and research works on the design framework and technical analysis of the initial TopicGraph model.

References

Bahdanau, D., Kyunghyun, C., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations.* San Diego, California, US.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research, Volume 3*, 993 - 1022.

Cho, K., Merriënboer, B. v., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics.

Chou, K.-H. (2020). A Method Combining Topic and GNN Models for Text Classification.

Chou, K.-H., Wu, Y.-T., & Wang, J.-H. (2020). Combining Dependency Parser and GNN Models for Text Classification. *Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)* (pp. 50–58). Taipei, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks.* Springer.

Hamilton, W., Ying, R., & Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems 30 (NIPS 2017).* Long Beach, California, US.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746–1751). Doha, Qatar: Association for Computational Linguistics.

Kipf, T., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. *International Conference on Learning Representations.* Toulon, France.

Liu, P., Qiu, X., & Huang, X. (2016). Recurrent Neural Network for Text Classification with Multi-Task Learning. *25th International Joint Conference on Artificial Intelligence IJCAI-16.* New York.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations.* Scottsdale, AZ, USA.

Schuster, M., & Paliwal, K. (1997). Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing (Volume: 45, Issue: 11, November 1997)*, 2673-2681.

Shen, D., Wang, G., Wang, W., Min, M., Su, Q., Zhang, Y., . . . Carin, L. (2018). Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 440–450). Melbourne, Australia: Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., . . . Polosukhin, I. (2017). Attention Is All You Need. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000-6010.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. *International Conference on Learning Representations.* Vancouver, BC, Canada.

Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., . . . Zhang, Z. (2019). Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. *ArXiv.*

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. (2021). A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions* *on Neural Networks and Learning Systems (Volume: 32, Issue: 1, January 2021)*, 4-24.

Yao, L., Mao, C., & Luo, Y. (2019). Graph Convolutional Networks for Text Classification. *33rd AAAI Conference on Artificial Intelligence (AAAI 2019)*, (pp. 7370-7377). Honolulu, Hawaii, US.

Zeng, H., Zhou, H., Srivastava, A., Kannan, R., & Prasanna, V. (2020). GraphSAINT: Graph Sampling Based Inductive Learning Method. *International Conference on Learning Representations.* Online.

基於時序模型和圖神經網路之 NBA 季後賽勝負預測
Predicting the Outcome of NBA Playoffs Based on Time Series Model and Graph Neural Network

Yen-Tsang Wu　**Jenq-Haur Wang**
Department of Computer Science and Information Engineering
National Taipei University of Technology
Web Information Retrieval Lab
buddyswu@gmail.com
jhwang@ntut.edu.tw

Ning Chien
Chunghwa Telecom Laboratories Advanced Technology Laboratory
nickchien@cht.com.tw

摘要

近年來，關於賽事結果預測的研究普遍存在三個問題。第一，許多研究使用比賽結束後的球員統計數據來預測該場比賽的結果；第二，這些研究通常基於球隊的平均表現進行分析和預測。此外，對於球賽數據預測的方式多採用傳統的統計模型，這樣的作法並未考量到數據之間的相關性。以上問題導致了賽事預測的效能偏低。本文提出了一種基於時序模型與圖神經網路的架構，用於預測季後賽球隊的勝負結果。首先，我們將球員作為圖中的節點（nodes），並將時序模型預測的球員表現作為節點特徵（node features），根據球員在球隊中的位置關係建立邊（edges），構建圖（graph）。其次，利用我們提出的圖神經網路架構進行預測，其中 GAT 的注意力機制（attention）用於選取圖中重要的節點並計算節點表達式（node representation）。隨後，經由 GCN 進行卷積（convolution）獲得特徵向量，並通過全連接層（fully connected layer）將節點表達轉換為圖表達（graph representation），最後進行賽事勝負預測。我們使用了美國職籃（National Basketball Association, NBA）2020-2021 球季的數據進行實驗。實驗結果顯示，使用我們提出的方法進行賽事結果預測，準確率達到 76.9%。結果表明，所提出的架構能夠有效預測比賽的勝負。

Abstract

In recent years, most studies on predicting game outcomes face two major issues. The first is the use of player statistics recorded after the game has ended to predict the outcome of that same game. The second issue is analyzing and predicting based on the average performance of the team. This paper proposes a novel approach for predicting game outcomes by integrating time series models and graph neural networks (GNNs). First, player performance predicted by time series models is treated as node features, and edges are constructed based on the players' positional relationships to form a graph. We then introduce a graph neural network architecture for prediction, where the attention mechanism of GAT is used to select important nodes in the graph and compute their representations, while GCN is employed to perform convolution to derive feature vectors. Finally, the node representations are transformed into graph representations for predicting the final game outcomes. The prediction accuracy of the proposed method reaches 76.9%, demonstrating its effectiveness.

關鍵字：選手表現預測、NBA 賽事勝負預測、圖神經網路、機器學習

Keywords: Player Performance Prediction, Graph Neural Networks, Machine Learning,

1 Introduction

隨著運動經濟的蓬勃發展，NBA 已成為全球增長最快的產業之一。然而，現有的比賽勝負預測多依賴賽後數據，忽略了賽前數據的重要性。此外，目前研究主要基於球隊的平均數據進行預測，缺乏對個別球員表現的考

量。因此，現有研究在球隊勝負預測上面臨兩大問題：(1) 賽前對球員表現的預測不足；(2) 缺乏對球員間互動的考慮。為了解決上述問題，我們提出了一種基於深度學習的 Gated Recurrent Unit (GRU) 和圖神經網路 (GNN) 的方法，以提高預測的準確性。我們的研究貢獻如下：

1. 在數據預測方面，我們使用時序模型來預測賽前球員表現，取代傳統的數學運算方式，使賽事勝負預測更貼近真實情況。
2. 本研究考量了所有上場球員的表現，並將圖神經網路應用於賽事預測，預測準確率達到 76.9%，超越了現有的研究方法。

2 Related Work

由於 NBA 賽事擁有豐富的統計數據，因此吸引了大量相關研究。Greene(Greene, 2015)利用綜合數據模型來估算大學新秀的 NBA 選秀順位，這項研究突顯了通過詳細的統計分析和數據建模來增強對 NBA 選秀中球員潛力的理解。Hu 等人(Hu et al., 2019)則透過神經網路預測 NBA 年度最有價值球員（MVP）的人選。Sarlis 等人(Sarlis et al., 2021)利用深度學習的方法來評估傷病如何影響球員的個人表現和整體球隊的戰績。Farghaly 等人(Farghaly & Deshpande, 2024) 則使用多種機器學習的方法探討 NBA 球員下肢肌肉拉傷的可能性，並分析了不同因素對受傷風險的影響。現今有許多的研究運用空間和時間的資料，結合深度學習、集成學習及統計方法，對季後賽結果、球員受傷與否和選秀順位進行預測。由於本研究的主題聚焦於 NBA 賽事結果預測，因此相關文獻主要集中在三個方面：籃球比賽結果預測、特徵選取，以及圖神經網路（GNN）在運動賽事中的應用。

2.1 Basketball Game Outcome Prediction

賽事結果預測可以視為一種二元分類任務。Hu 等人(Hu & Zidek, 2004)利用數值分析方法，通過統計主場和客場的勝負差異，並使用 weighted likelihood 計算 1996-1997 年冠軍賽的勝負結果。Miljković 等人(Miljković et al., 2010) 將比賽數據分為主場和客場，並使用 Naïve Bayes 和 multivariate linear regression 作為分類模型，預測準確率達到 67%。Cao 等人(Cao, 2012) 使用 2005-2010 賽季作為訓練數據，以 2010-2011 賽季為測試數據，採用 Naïve Bayes、Support Vector Machine (SVM)、Logistic Regression 等模型進行預測，準確率達到 69.67%。

Pai 等人(Pai et al., 2017)提出一個 HSVMDT 的架構，結合 SVM 和 Decision Tree，預測準確率達到 85.2%。Jain 等人(Jain & Kaur, 2017)提出了 Hybrid Fuzzy-SVM (HFSVM）模型，用於降低數據中 noises 對 SVM 的影響，減少預測誤差，並且使用 CFS（Correlation-based Feature Selection, Hall 1999）技術進行特徵選取，預測準確率達到 88.26%。Horvat 等人(Horvat et al., 2020)使用多個賽季的數據進行訓練，並採用 Decision Tree、KNN 演算法和 Random Forest(RF)進行實驗，預測準確率達到 60.8%。Osken 等人(Osken & Onay, 2022)使用 K-Means 和 C-Means 聚類演算法來識別球員類型，並基於成員的能力訓練預測模型，他們的方法在 NBA 的五個賽季中實現了 76% 的預測準確率。Wang 等人(Wang, 2023) 則是進行了特徵工程，透過分析和選擇關鍵指標來提高模型預測準確性，研究發現數據中，投籃命中率、三分球命中率和搶斷數是預測比賽結果的關鍵指標。他們使用 Random Forest 和深度神經網路（DNN）來進行預測，最佳準確率達到 74%。Adam 等人(Adam et al., 2024) 利用比賽上半場的統計數據，並使用 SVM 來預測最後勝負結果，預測準確率達到 66.67%。

2.2 GNNs Methodology Used in Sports

圖神經網絡（Graph Neural Networks, GNNs）是一種用於處理圖結構數據的神經網路模型，通過利用圖中節點（nodes）與邊（edges）之間的關聯性和相互依賴性進行信息傳遞與學習。由於體育比賽中的數據是高度多樣化的，包括時間序列、空間位置以及球員之間的互動。GNN 能夠將這些多模態數據有效地整合，處理體育運動中的圖結構數據，捕捉球員、球隊和比賽之間的複雜關聯，從而提供更準確和有效的預測與分析。

Xenopoulos 等人 (Xenopoulos & Silva, 2021) 使用 GNN 預測美式足球（NFL）和電子競技遊戲（CSGO）的比賽結果，分別將損失（LOSS）減少了 9% 和 20%。這表明 GNN 在

處理這類高互動性和多維數據方面具有顯著優勢。Zhao 等人(Zhao et al., 2023) 的研究結合了 GCN 與 Random Forest 算法，以提升對 NBA 比賽結果的預測準確性，預測準確率達到了 71.54%。Luo 等人 (Luo & Krishnamurthy, 2023) 則提出了一種名為 GATv2-TCN 的深度學習方法，將圖注意力網路（Graph Attention Networks, GAT）與時間卷積網路（Temporal Convolutional Networks, TCN）結合，用於預測運動表現。他們的研究探討了如何通過這些模型捕捉時間序列與圖結構之間的動態關聯，從而提高預測精度。

2.3 Prediction Based on Feature Selection

由於 NBA 數據屬性繁多，特徵選取（feature selection）在比賽勝負預測中至關重要。Thabtah 等人(Thabtah et al., 2019b)除了傳統數據以外，更加入了主場（home）和客場（road）因素，分別使用 Multiple Regression(Berger, 2003)、CFS(Hall, 1999)和 RIPPER algorithm(Cohen, 1995)等方法挑選特徵。研究將這些方法選取的特徵分成五組進行實驗，發現使用 RIPPER algorithm 與 Naïve Bayes 結合的模型表現最佳，準確率達 83%。另外，作者在實驗中還發現這五組特徵中都有一個共同的特徵，即「DRB」(Defensive Rebounds, 防守籃板），因此推論這是對 NBA 賽事勝負預測中最重要的一項特徵。以上研究都表明，特徵選取在提升比賽預測效能中扮演了至關重要的角色，對比賽結果預測有顯著影響。

3 Methodology

本研究所提出的方法主要分為兩個部分：第一是**球員表現預測（Player Performance Prediction）**：由於我們無法在比賽開始前就獲得球員的表現數據，因此，我們需要透過球員過去的表現預測出該場比賽的球員數據。第二是**季後賽勝負預測（Game Outcomes Prediction）**：在這一部分，我們使用圖神經網路（GNN）來進行勝負預測。如前面章節所提，輸入的特徵至為關鍵，因此在進行預測前，通過特徵選取（feature selection）來找出球員表現中對比賽結果最具影響的關鍵特徵。這些特徵將被用作 GNN 的輸入，進行賽事結果的預測。系統流程如圖 1 所示。

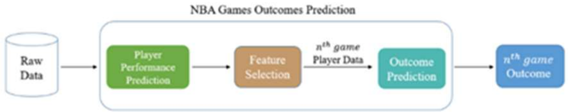


圖 1. 系統流程圖

我們的方法主要分為以下三個步驟：球員表現預測、特徵選取和比賽勝負預測。

3.1 Player Performance Prediction Model

在比賽尚未開打前，我們並無法得知球員的表現，因此，對球員表現的準確預測成為預測比賽結果的關鍵因素。透過觀察，我們可以發現球員的表現往往具有維持一段時間的趨勢，所以我們將預測球員表現當作時序特徵預測任務。

首先，我們將所有數據進行標準化（normalization），接著使用基於時序數據的 many-to-one GRU(Dey & Salem, 2017)來預測球員表現。該模型以多場比賽的數據作為輸入，輸出預測單一特徵的值，用來推測下一場比賽中每位球員的表現。GRU 模型的 input 為球員過去的表現數據，輸入的數據必須轉換成(n, m, k)的形式才能進行訓練，其中 n 代表輸入的比賽場次(batch size)，m 為 sliding window 的大小，k 為球員表現的特徵數量，將轉換好的球員數據輸入模型進行訓練，最後經過一個線性層(Linear Layer)將模型的輸出層向量轉換成實數空間，進行預測。最終，我們可以得到第 n +1 場比賽中每位球員的表現預測。由於總共有 21 項不同的球員表現特徵，因此我們需要執行 21 次預測來完成一場比賽完整的球員數據預測。提出的整體架構如圖 2 所示。

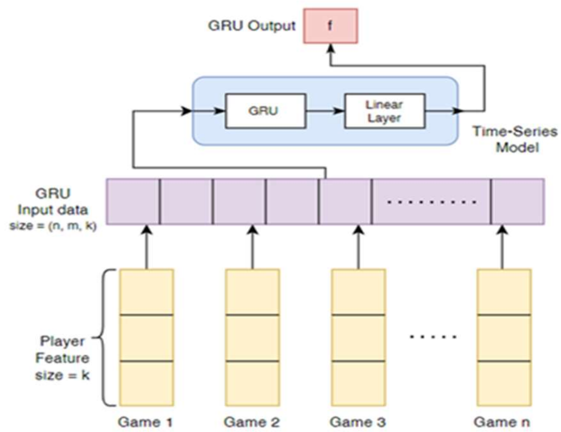


圖 2. 選手表現預測架構圖

我們每次只預測一種特徵，並使用 Sliding Window 的方式，將前 n 場比賽當作訓練來預測第 n+1 場比賽中球員的表現。圖 3 為利用前 3 場（即g_1, g_2, g_3）的球員表現預測第 4 場(即

g_4)的球員表現的例子。其中g_i代表第 i 場比賽，而i∈[1,n]。

圖 3 sliding window 示意圖

我們使用 Mean-Square Error(MSE)來評估時序模型，即預測值與實際值之間距離的平方和。

3.2 Feature Selection

NBA 官方公布了豐富的球員數據，這些數據被分為傳統數據和進階數據。傳統數據是球員在比賽中的真實表現，例如得分、籃板、助攻等具體統計項目，這些數據可以直接反映球員在場上的實際表現，因此能夠被時序模型用來進行預測。進階數據則是根據傳統數據進行進一步計算和推導而得出，例如使用命中率、效率值等複合指標來評估球員的表現。本研究採用傳統數據進行實驗，因為它直接反映了球員的比賽表現，詳細說明如表 1.

英文縮寫	中文名稱	英文縮寫	中文名稱
MP	上場時間	TS%	真實命中率
FG	命中球數	eFG%	有效命中率
FGA	總出手球數	FTr	罰球製造率
FG%	命中率	ORB%	進攻籃板率
TP	三分進球數	DRB%	防守籃板率
TPA	三分總數	TRB%	總籃板率
TP%	三分命中率	AST%	助攻率
FT	罰球進球數	STL%	抄截率
FTA	罰球總數	BLK%	阻攻率
FT%	罰球命中率	TOV%	失誤率
ORB	進攻籃板	USG%	球權佔有率
DRB	防守籃板	ORtg	進攻率
TRB	總籃板	DRtg	防守率
AST	助攻	TOV	失誤
STL	抄截	PF	犯規次數
BLK	阻攻	PTS	得分
+/-	場上效率值		

表 1. 數據列表

球員特徵的選取將在數據預測完成後進行篩選，目的是確認每項球員特徵對於賽事勝負預測的重要性。本文使用 SVM 來進行特徵選取。

3.3 Playoff Outcomes Prediction

籃球是一種團隊運動，所有球員彼此相互影響，因此，為了找出影響勝負的關鍵球員以及球員之間的互動影響，本文提出圖注意力卷積網路(Graph Attention Convolution Neural Network, GATCN)，一種結合 GAT(Veličković et al., 2017b)和 GCN(Li et al., 2018)的新架構以進行賽事預測，如圖 4 所示:

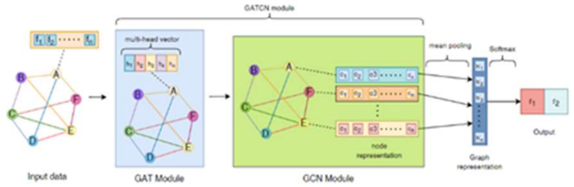


圖 4. GATCN 架構圖

我們首先將球員表現轉換成圖結構形式，並將其作為 GATCN 模型的 input。接著，通過 GAT 計算每個節點（球員）與其 1-hop 鄰居之間的特徵向量。GAT 能夠根據節點之間的重要性分配注意力係數，並使用這些係數來更新每個節點的特徵向量，然後，透過 GCN 將節點之間的區域連結和全域連結性進行聚合，以得出新的節點表達式。接著，通過 mean pooling，將更新後的節點表達式轉換成整體的圖表達式，最後，經由 softmax 函數，我們得到最終 output，即該場比賽的勝負預測。

3.3.1 Player-to-Player Graph

為了能夠精準計算每位球員對比賽的影響值，我們提出了一種 Player-to-Player Graph 的建圖方式。我們將每場比賽中上場的所有球員作為節點，根據他們之間的關聯性來構建出一張圖。這種建圖方式能夠將球員的互動特徵轉換為圖結構，我們更能夠有效捕捉球員之間的複雜關聯性和互動模式，反映球員的團隊合作以及個別球員在比賽中的影響力。

首先，給定一張圖G=(V,E)，G 代表一場比賽，V 為球員 p 集合，表示為 $V=\{p_1,p_2,\dots,p_n\}, n\in\mathbb{R}$，其中 n 為一支球隊有上場的球員數量，邊$E=\{e_1,e_2,\dots,e_n\}, n\in\mathbb{R}$ 則為球員跟球員間的關係，由於 NBA 官方並未公布每位先發和替補球員的上場時間，因此本研究會以同隊的球員建邊形成全連接圖，如圖 5 所示。

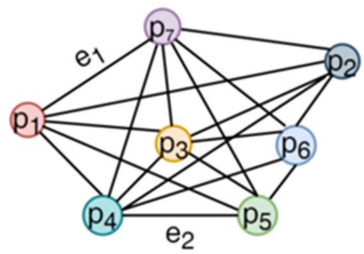


圖 5.Player-to-Player graph 示意圖

球員之間的邊會以相鄰矩陣 A 表示，對於每個相鄰矩陣元素A_{ij}代表球員 p_i 和 p_j之間的邊，其中$i,j\in\mathbb{R}$。我們將同隊的球員之間建邊，如公式(1)所示。若p_i和p_j是同隊的關係，則相鄰矩陣A_{ij}的值為 1，若不是同隊的關係則A_{ij}

為 0，即 $A_{ij} = A_{ji}$，形成無向圖(undirected graph)，節點的特徵（node feature）是該球員在該場比賽中的表現數據。

$$f(x) = \begin{cases} 1, & if\ (p_i, p_j) \in E\ and\ i \neq j \\ 0, & others \end{cases} \quad (1)$$

3.3.2 Graph Attention Convolution Network

我們提出的 GATCN 結合了 GAT 和 GCN 所運用到的技術，GATCN 在 propagation 過程中會運用到自注意力機制(self-attention)，該機制通過計算每個節點與其鄰居節點之間的 注意力係數，來判斷節點的重要性。如公式 2 所示。

$$\alpha_{ij} = \frac{\exp\left(LeakyReLU\left(\vec{a}^T[\mathbf{W}\vec{h}_i \parallel \mathbf{W}\vec{h}_j]\right)\right)}{\sum_{k\in N_i}\exp\left(LeakyReLU\left(\vec{a}^T[\mathbf{W}\vec{h}_i \parallel \mathbf{W}\vec{h}_k]\right)\right)} \quad (2)$$

α_{ij} 為節點(i , j)的注意力係數，$i , j \in \mathbb{R}$，其中 $\vec{h}, \vec{a}$ 的定義如之前所述，以 LeakyReLU(Xu et al., 2015)當作啟動函數調整權重，我們運用 multi-head self-attention，讓各個 head 能夠關注更多區域和全域的訊息，為了將所有 head 得出的特徵向量結合，形成 multi-head，我們透過公式 3.的計算，將所有向量做串聯。

$$h_i' = \parallel_{k=1}^{K} \sigma\left(\sum_{j\in N_i} \alpha_{ij}^k W^k h_j\right) \quad (3)$$

$\parallel$為串聯符號，K 為 head 數量，σ為 sigmoid 啟動函數，α_{ij}^k為節點(i,j)在第 k 個 head 的注意力係數，W^k為第 k 個 head 的線性轉換權重矩陣，h_i'為 K 個 head 串連起來的新注意係數，代表各個球員對於一場比賽的權重。Graph 具有區域連接(local connected)的特性，我們利用卷積的方式來聚合特徵，由於此階段的特徵已經經由 GAT 計算過，因此包含和 1-hop 鄰居之間的注意力權重，接著透過 GCN 聚合所有鄰居和自身節點的特徵訊息，更新後的點特徵向量就具備全域信息。為了進行勝負預測，我們沿用 Monti 等人(Federico Monti & Bronstein, 2019)對於將點特徵轉換成圖特徵的方法，透過 mean pooling 的方式將每個 batch 的點特徵向量轉換成圖特徵向量，如公式(4)所示，

$$x_i = \frac{1}{N_i}\sum_{n=1}^{N_i} h_i' \quad (4)$$

N_i為節點 i 鄰居節點數量，x_i為圖特徵向量，代表比賽中所有球員的資訊，最後再經過 softmax 的轉換，會得出一個二維向量，代表的意義為比賽勝利或失敗的機率。在訓練的過程中以交叉熵(cross-entropy)作為損失函數。為避免 over fitting 的狀況發生，我們會計算出每一次訓練過程的 loss 值H_i，其公式如(5)所示，其中 C 為類別數，在此為勝負兩種維度，n 為所有資料的筆數，y_c為類別做 one-hot encode 後的第 i 筆數據的標籤，$r_{c,i}$為第 i 筆資料屬於 c 類的機率

$$\mathrm{H}_i = \sum_{c=i}^{C}\sum_{i=1}^{n} -y_c \log_2(r_{c,i}) \quad (5)$$

4 Experiment and Results

在這項研究中，我們的目標是準確預測 NBA 季後賽的結果。接下來的章節將詳細說明我們所使用的數據集，實驗程序和結果。

4.1 NBA Dataset

本研究實驗的資料集是從 www.basketball-reference.com 中，通過網路爬蟲抓取 NBA 2020-2021 賽季的公開數據而來。該賽季因為受到新冠肺炎(COVID-19)影響，每支球隊只進行 72 場例行賽，因此例行賽總數共有 1080 場，而季後賽則有 91 場。每場比賽有兩支球隊參與，因此每場比賽會產生出一勝一負的數據，因此總數為 2160 場。資料集包含了每場比賽中球員的傳統數據、進階數據、球隊平均數據、勝負(label)。為了避免重複預測，我們採用主場球員表現做為預測對象，訓練集和驗證集則依照 8:2 的比例切割，詳細數據如表 2 所示:

	訓練集	驗證集	測試集
勝	509	127	53
敗	423	107	38
總數	932	234	91

表 2. 資料集統計表

4.2 Evaluation Metrics

我們以準確率(accuracy)、精確率(precision)、召回率(recall)和 f1-score 進行模型的效能評估。

4.3 Player Parameter

由於我們的模型是基於球員表現進行勝負預測，因此每場參與比賽球員的數量會是預測的重要參數之一，我們透過統計 2016-2019 賽季共 681 場的季後賽上場人數作為參考，我們發現板凳球員上場時間平均在 15-20 分鐘之間。根據 NBA 統計顯示，每場比賽只使用 7 個球員的場次最多，達到 280 場，其次是使用 8 名球員的比賽場次，共有 203 場。因此，我們將模型的球員數量設為 7 跟 8 人，進行實驗比較。

4.4 Compare models

為了驗證提出模型的效能，我們使用以下模型作為比較對象：

- Baseline：我們基於兩篇論文的方法進行融合，以產生一個基準模型。首先，根據 Jones(Jones, 2016)提出 3-game-average 的方法，預測出球員表現後，再使用 XGBOOST 做特徵選取，計算平均成為隊伍整體數據。接著我們使用 Thabtah (Thabtah et al., 2019a)等人提出的 ANN 模型進行比賽結果預測。
- GAT：由 Veličković等人(Veličković et al., 2017a)所提出的架構，將圖神經網路加上注意力機制，使得模型能夠關注單一個點和其 1-hop 鄰居，並計算出注意力係數更新點特徵向量。
- GCN：由 Monti 等人(Federico Monti & Bronstein, 2019)提出一種基於 GCN 的圖分類模型，並運用於真假新聞判斷上。其架構的核心為使用兩層的卷積層與全連接層，將 GCN 輸出的 64 維向量降至 2 維，並透過 softmax 計算真假類別的機率。
- DGCNN：由 Zhang 等人 (Zhang et al.)提出的 GCN 改良模型，總共有四層 GCN，透過 sort pooling 的方法將每一層的輸出，藉由節點在圖中的結構角色(structural roles)進行排序，然後將 pooling 完的結果輸入傳統 1-D convolution 進行卷積，最後經過線性層輸出結果。

4.5 Experimental Results

如表 3 所示，我們提出的 GATCN 模型，在各項評估指標上都有優異的表現。

Model	Accuracy	Precision	Recall	F1-score
Baseline	0.593	0.886	0.602	0.717
SVM+GAT	0.736	0.754	0.784	0.769
SVM+GCN	0.736	0.792	0.763	0.777
SVM+ DGCNN	0.747	0.773	0.788	0.781
GATCN	0.652	0.769	0.666	0.714
SVM+ GATCN	**0.769**	0.811	**0.796**	**0.803**

表 3. 實驗結果總表

當我們使用 sliding window = 3 作為 GRU 的設置對球員的表現進行預測，然後按照每場比賽的 8 位球員建圖。接著，通過使用 SVM 挑選出的 8 個特徵作為 GATCN 的輸入進行訓練時，我們的方法在準確率、召回率和 F1-score 的表現都是最好，分別達到 0.769、0.811 和 0.803。實驗結果證明，經過特徵選取後的 GATCN 模型在預測球隊勝負時有較好的表現。我們提出的模型以 GAT 聚合(aggregate)每個節點的 one-hop 鄰居，透過加權平均更新節點特徵，再使用 GCN 進行 Laplace matrix 運算將所有鄰居的特徵聚合，比起只單純使用 GAT 或 GCN 的模型，GATCN 在預測準確率上分別贏過 GAT 和 GCN 模型 3.3%。這充分證明了 GATCN 能夠有效融合兩者的優點，從而提高模型性能。我們也觀察到，DGCNN 模型在 F1-score 有 0.781 的優異表現。DGCNN 通過多層的 GCN 計算出節點與鄰居的關係，並使用 sort pooling 來對節點特徵先排序再 pooling 的方法，對於提升 GCN 是有效益的，但由於缺少了注意力係數的幫助，在準確率仍低於 GATCN 模型 2%。

5 Analysis and Discussion

在這個章節，我們對球員的特徵選取，時序模型以及其他的研究進行討論與分析。

5.1 Analysis of Feature Selection

在構建球員表現預測模型時，挑選哪些數據作為球員的特徵是影響模型性能的關鍵因素。為了選找出最重要的特徵，我們通過 SVM 計算特徵的重要性來進行特徵選取。

如圖 6 所示，SVM 計算出的特徵值範圍從 -1 到 1，趨近於 1 的特徵表示該特徵與比賽勝利正相關；而趨近於 -1 的特徵則表示該特徵與比賽勝利負相關。根據這些計算結果，我們挑選出出[TRB, STL, PTS, FG%, BLK, FT%,+/-,3P%, ORB, DRB, FGA, TOV, FG, FTA, 3P, FT]作為一場比賽的特徵。其中 TRB 的數值是最大的，代表總籃板和球隊贏球最為關鍵；而 ORB 的數值是最小的，代表進攻籃板和球隊輸球的關係最相關，而[MP, AST, PF, 3PA]則因數值都趨近於 0，表示無關於比賽勝利和失敗。

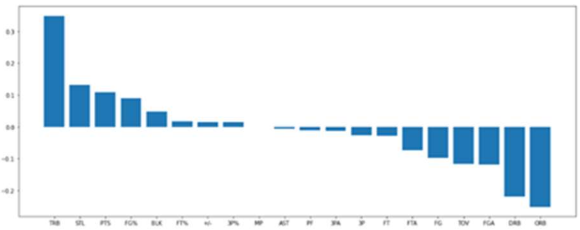


圖 6. SVM 特徵分數

在過去的研究中，常見的特徵選取方法包括卡方檢驗（Chi-Square）、Random Forest 和 XGBoost 等。這些方法通常會通過計算特徵的分數來篩選出對模型最有用的特徵。根據的 Chen 等人(Chen et al., 2021)的研究，本研究選擇特徵分數高於平均值的特徵進行比較，以便挑選出最具影響力的特徵。

除了球員的個人數據之外，建邊的球員的數量也會是重要的參數，正如前述章節所提到，7 到 8 名球員是比賽中最常使用的人數。因此，我們在實驗中選擇 7 和 8 人的球員數量來進行實驗。為了找出最佳的特徵，我們將以上四種特徵選取方法，使用由不同 sliding windows 大小產生的球員數據以及不同的選手數量與 GATCN 結合，驗證其效果。其中 sliding windows 以場數表示，實驗結果如下:

場數	人數	Method	Accuracy	Precision	Recall	F1-score
2	8	Baseline	0.652	0.692	0.692	0.692
2	8	SVM	0.681	0.717	0.730	0.723
2	8	XGBOOST	0.637	0.679	0.692	0.685
2	8	RF	0.582	1.000	0.582	0736
2	8	Chi-square	0.593	0.962	0.593	0.733
3	8	Baseline	0.695	0.769	0.714	0.740
3	**8**	**SVM**	**0.769**	**0.811**	**0.796**	**0.803**
3	8	XGBOOST	0.758	0.811	0.781	0.796
3	8	RF	0.681	0.660	0.760	0.707
3	8	Chi-square	0.692	0.735	0.735	0.735
4	8	Baseline	0.565	1.000	0.565	0.722
4	8	SVM	0.637	0.679	0.692	0.685
4	8	XGBOOST	0.626	0.660	0.686	0.673
4	8	RF	0.582	1.000	0.582	0736
4	8	Chi-square	0.593	0.962	0.593	0.733

表 4. Feature selection 之實驗結果

從表 4 可以看出，當使用 8 位球員建圖並且以 3 場比賽作為 sliding windows 的場數時，模型在準確率、精確率、召回率和 F1-score 上的表現優於使用 2 場或 4 場比賽，其中 4 場的效果是最差的。而 XGBoost 和 Random Forest 在 2 場的情況優於 4 場，這表明球員的表現並沒有長期的規律可供 GRU 進行學習。3 場比賽作為輸入是最好的。而以 SVM 作為特徵選取的方法時，模型的準確率、召回率和 F1-socre 分別是 0.769、0.796 和 0.803，是所有特徵選取的方法中最好的。而在精確率的表現上，使用 2 場和 4 場作為訓練的選手表現，並通過 Random Forest 進行特徵選取後達到 1。但觀察其結果發現是全部預測比賽勝利，因此造成精確率最高，但準確率卻是最低的 0.582。

在特徵選取方法的部分，使用 SVM 表現都是最好的，而使用 XGBOOST 則都是第二。進一步觀察，兩者分析出的特徵極為為相近，其中 TRB, STL, PTS, FG%, BLK, FT%, +/-, 3P% 等 8 項是相同的，因此它們在各項評估標準中的得分也相對接近。XGBOOST 是透過計算 gain 來評估特徵的重要度，減少模型的 entropy，使得重要信息能夠被放大，得出的特徵重要度更為精準，這是 XGBoost 在特徵選取方面的表現更優異的原因。相比之下，Random Forest 由於每次 生成的樹會因為 gini index 的不同會影響樹葉的分裂，導致 Random Forest 的特徵選取在此實驗中的表現不如其他方法。至於 Chi-Square，則更適合用於分析類別變數的相關性，但由於輸球和贏球的特徵在本研究中的差異不明顯，因此無法透過計算期望值的方式有效找出關鍵特徵，所以表現不如其他三種方法。

5.2 Analysis of Time Series Model

我們對著名的時序模型 GRU 和 LSTM 進行效能比較。實驗中，我們將兩個模型的場數參數設為 3，球員參數設為 8，並進行 50 epochs 的訓練，optimizer 則是使用 adam(Diederik, 2014)，Loss function 為 MSE。在兩者預測出球員表現後，我們將這些結果輸入 GATCN 模型，進行比賽勝負的預測。結果如表 5 所示:

Method	Accuracy	Precision	Recall	F1-score
LSTM	0.7033	0.7358	0.7500	0.742
GRU	0.7692	0.8113	0.7963	0.8037

表 5. GRU 和 LSTM 效果比較表

從實驗結果，我們可以發現 GRU 在第 15 epochs 時，所有特徵都已經達到收斂，而 LSTM 的 FT 則在第 20 epochs 才收斂。在使用 SVM 進行特徵選取後的實驗結果中，GRU 在四項評估指標上均優於 LSTM。這結果顯示 GRU 用更新閥取代 LSTM 的輸入閥和遺忘閥的作法，有顯著的提升效果。GRU 的優勢在於其更新閥設計，使得在計算單元狀態時能夠將隱藏層的資訊納入其中，從而傳遞了更多的全局信息。相比之下，LSTM 將隱藏層和單元狀態分開計算，導致它只能傳遞部分信息，因此造成了準確率和效能上的差異。此外，由於 GRU 模型結構更精簡，模型訓練的時間比 LSTM 縮短了 2.5%。

5.3 Analysis of GATCN

為了驗證我們所提之 GATCN 的效能，我們比較 Jain 等人(Jain & Kaur, 2017)和 Zhao (Zhao et al., 2023)等人提出的模型。由於 Jain 等人提出 HFSVM 的模型使用 NBA 2015-2016 賽季數據進行實驗，因此我們也使用相同的賽季數據

來測試我們的模型。實驗結果，GATCN 的準確率達到了 0.9302，顯著高於 HFSVM 的準確率為 0.8826。而與我們同樣基於圖神經網路的 Zhao 等人提出的模型，在同樣賽季中的最高準確率為 0.707。這意味著我們的模型對比 Jain 等人的模型準確率提升了 4.8%，而對比 Zhao 等人的模型則提升了 22%。由此可以說明在 NBA 勝負預測任務中，比起透過用團隊數據的模型相比，球員表現所提供的資訊能夠更精準的預測比賽的勝負。實驗準確率如圖 7 所示。

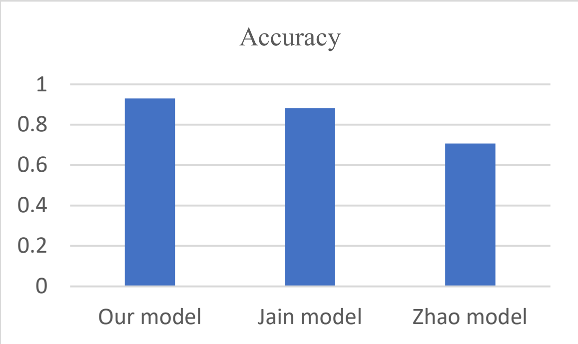


圖 7. 三個模型在 2015-16 賽季準確率比較圖

6 Conclusion and future

在本研究中，我們提出一個針對 NBA 季後賽的比賽結果預測方法。首先透過 GRU 預測選手未來的表現，並把所有球員當作 node，建立一個全連接圖。接著，通過 SVM 進行特徵選取，最後我們運用本文提出的圖注意力卷積網路（GATCN）進行預測。相較於僅使用球員表現數據進行預測，我們的方法更多了球員和球員之間的結構特徵，使得模型在賽事勝負預測中更加精確。

實驗結果顯示，本研究所提出的 GATCN 在準確率、召回率和 F1-score 達到 0.769、0.796 和 0.803，都超越了其他 state-of-the-art 的方法。

Acknowledgments

The authors would like to thank the supports from the National Science and Technology Council, Taiwan under the grant numbers: NSTC113-2221-E-027-096, and NSTC113-2634-F-027-001-MBK.

References

Adam, C., Pantatosakis, P., & Tsagris, M. (2024). On predicting an NBA game outcome from half-time statistics.

Berger, D. E. J. U. C. G. U. (2003). Introduction to multiple regression.

Cao, C. (2012). Sports data mining technology used in basketball outcome prediction.

Chen, W.-J., Jhou, M.-J., Lee, T.-S., & Lu, C.-J. J. E. (2021). Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association. *23*(4), 477.

Cohen, W. W. (1995). Fast effective rule induction. In *Machine learning proceedings 1995* (pp. 115-123). Elsevier.

Dey, R., & Salem, F. M. (2017). Gate-variants of gated recurrent unit (GRU) neural networks. 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS),

Diederik, P. K. J. (2014). Adam: A method for stochastic optimization.

Farghaly, O., & Deshpande, P. (2024). Leveraging Machine Learning to Predict National Basketball Association Player Injuries. 2024 IEEE International Workshop on Sport, Technology and Research (STAR),

Federico Monti , F. F., Davide Eynard, Damon Mannion,, & Bronstein, M. M. (2019). Fake News Detection on Social Media using Geometric Deep Learning. *ICLR*.

Greene, A. C. (2015). The Success of NBA Draft Picks: Can College Careers Predict NBA Winners?

Hall, M. A. (1999). *Correlation-based feature selection for machine learning* The University of Waikato].

Horvat, T., Havaš, L., & Srpak, D. J. S. (2020). The impact of selecting a validation method in machine learning on predicting basketball game outcomes. *12*(3), 431.

Hu, F., & Zidek, J. V. J. L. N.-M. S. (2004). Forecasting NBA basketball playoff outcomes using the weighted likelihood. 385-395.

Hu, J., Zhang, H., & Qiu, J. (2019). Prediction of MVP attribution in NBA regular match based on BP neural network model. Proceedings of the 2019 international conference on artificial intelligence and advanced manufacturing,

Jain, S., & Kaur, H. (2017). Machine learning approaches to predict basketball game outcome. 2017 3rd international conference on advances in computing, communication & automation (ICACCA)(Fall),

Jones, E. S. (2016). Predicting Outcomes of NBA Basketball Games.

Li, Q., Han, Z., & Wu, X.-M. (2018). Deeper insights into graph convolutional networks for semi-supervised learning. Proceedings of the AAAI conference on artificial intelligence,

Luo, R., & Krishnamurthy, V. J. a. p. a. (2023). Who You Play Affects How You Play: Predicting Sports Performance Using Graph Attention Networks With Temporal Convolution.

Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010). The use of data mining for basketball matches outcomes prediction. IEEE 8th international symposium on intelligent systems and informatics,

Osken, C., & Onay, C. J. H. (2022). Predicting the winning team in basketball: A novel approach. *8*(12).

Pai, P.-F., ChangLiao, L.-H., Lin, K.-P. J. N. C., & Applications. (2017). Analyzing basketball games by a support vector machines with decision tree model. *28*, 4159-4167.

Sarlis, V., Chatziilias, V., Tjortjis, C., & Mandalidis, D. J. I. S. (2021). A data science approach analysing the impact of injuries on basketball player and team performance. *99*, 101750.

Thabtah, F., Zhang, L., & Abdelhamid, N. (2019a). NBA Game Result Prediction Using Feature Analysis and Machine Learning. *Annals of Data Science*, *6*(1), 103-116.

Thabtah, F., Zhang, L., & Abdelhamid, N. J. A. o. D. S. (2019b). NBA game result prediction using feature analysis and machine learning. *6*(1), 103-116.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017a). Graph attention networks. *ICLR*

arXiv preprint arXiv:1710.10903.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. J. a. p. a. (2017b). Graph attention networks.

Wang, J. (2023). Predictive Analysis of NBA Game Outcomes through Machine Learning. Proceedings of the 6th International Conference on Machine Learning and Machine Intelligence,

Xenopoulos, P., & Silva, C. (2021). Graph neural networks to predict sports outcomes. 2021 IEEE International Conference on Big Data (Big Data),

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Zhang, M., Cui, Z., Neumann, M., & Chen, Y. (2018). An end-to-end deep learning architecture for graph classification. Proceedings of the AAAI conference on artificial intelligence,

Zhao, K., Du, C., & Tan, G. J. E. (2023). Enhancing basketball game outcome prediction through fused graph convolutional networks and random forest algorithm. *25*(5), 765.

中文文章級別人物關係擷取之研究
Research on Document-Level Person Relation Extraction in Chinese

洪閔昭
國立中央大學
111522155@cc.ncu.edu.tw

張嘉惠
國立中央大學
chia@csie.ncu.edu.tw

葉季儒
國立中央大學
tobyeider.ncu@g.ncu.edu.tw

摘要

本研究旨在構建一套可應用於眞實網路資料的聯合實體關係擷取架構。針對現有資料集來源單一且主要集中在句子級別的問題，我們利用大型語言模型（如 Gemini、GPT-3.5）對全篇文章進行標記，並使用中文 Common Crawl 數據構建更泛用的資料集。爲提高標記的可信度與實體對取樣的完整性，採用了交叉驗證與實體擴充方法。並通過微調預訓練模型來驗證與提升模型在眞實環境下進行實體關係擷取的性能。

關鍵字：命名實體識別、聯合實體關係擷取、文章級關係擷取

1 Introduction

關係擷取（Relation Extraction，簡稱 RE），旨在從文本中識別出實體間的關係。在傳統的關係擷取任務上，通常需要先進行命名實體識別 (Named Entity Recognition，簡稱 NER）任務，即從文本中識別出具有指稱性的實體，如人物、地點、組織等。其次，判斷兩個實體間是否存在關係，最後再進行關係分類。而命名實體識別以及關係擷取兩個任務合在一起則稱爲聯合實體關係擷取 (Joint Entity and Relation Extraction)

近年來由於大型語言模型 (LLM) 的快速發展，聯合實體關係擷取的任務逐漸改由生成式模型來解決，而在先前的研究中也表明了，生成式模型可以在關係擷取達到 State-of-the-art (SOTA) 的效果 [1]。且在 [2] 的實驗中也發現像 GPT-3 規模的模型，即便不需經過 fine-tune，只需要給定良好的 instruction，即可達到 SOTA 的成效。

不過現今的關係擷取及命名實體類別任務資料集，通常來源都是單一特定的資料庫，例如: 維基百科、新聞網站，因此，訓練出來的模型較難泛化到眞實世界中多樣的網路數據上。再者，目前在關係擷取任務的主要資料集，例如:ACE05[3]、CoNLL[4]、NYT[5] 等資料集，都是屬於 Sentence 級別的任務，模型只需要在較短的語句或段落中找出實體以及關係即可。但在現實的情況下，實體關係三元組中的實體以及關係，並不一定會同時出現在同一個句子，或是相鄰的句子中。根據 Yao 等人 [6] 統計，超過 40.7%的關係只能在文章級別上被識別出來。而少數像 DocRED[6] 這類的文章級別資料集，目前也都是以英文的資源爲主，缺乏相對應的中文資源。

在本篇論文中，我們利用當前最先進的大型語言模型（如 Gemini、GPT-3.5 等），對 Common Crawl 全網爬蟲資料內容進行標記，創建中文文章級別關係擷取資料集。有別於傳統句子級別的關係擷取任務，我們的方法讓模型在整篇文章中進行跨句子、跨段落的關係擷取，也順利找出了大約 30% 只能在文章級別上被識別出來的關係。這克服了過去因文章長度限制而必須截斷文本或進行證據檢索的局限。而我們也在參數量較小的模型中實驗證實，該資料集所訓練出的模型，可泛化於眞實網路文章中的模型可行性，並爲未來的相關研究留下了基線做爲參考。

2 Related Work

關係擷取（RE）是自然語言處理（NLP）中的關鍵任務，廣泛應用於知識圖譜構建 [7, 8]、問答系統 [9, 10, 11]、對話系統 [12, 13] 等領域。傳統的 RE 方法將其視爲多類別分類問題，通過 LSTM 或 BERT 等模型進行監督訓練 [14, 15]，但這些方法依賴大量人工標記數據。爲減輕標記負擔，一些研究引入遠程監督 [16] 和半監督方法 [17]，但這可能導致不準確的標記。

傳統 RE 多使用 pipeline 方式，先抽取實體再進行關係分類，但這忽略了命名實體識別（NER）和關係擷取（RE）間的互動，導致誤差傳遞問題。爲改善此問題，常用多任務學習 [18, 19] 和更換擷取策略的方式，如 CasRel[20]、ETL-Span[21] 等，通過共享參數

或改變擷取步驟來減少誤差。然而，這些方法仍受限於誤差傳遞和暴露偏差問題 [22]。

爲解決這些問題，研究者提出了填表方法，例如 TPLinker[23] 和 UNIRE[24]，通過維護關係表來擷取三元組，避免了誤差傳遞和暴露偏差問題，但在處理長文本時效率較低。

2.1 生成式聯合實體關係擷取

自 2018 年生成式聯合實體關係擷取方法如 CopyRE[25] 提出以來，研究逐漸轉向使用預訓練模型（如 UniLM[26]、Bart[27]、T5[28]）進行聯合實體關係擷取。例如，Ye 等人 [29] 提出基於 UniLM 的 CGT 對比學習方法，Cabot 等人 [30] 則以 BART 爲基礎提出 REBEL 架構。

生成式擷取可分爲多輪式生成和通用式生成。多輪式生成將實體和關係的擷取視爲多輪問答對話的問題，如 Li 等人 [31] 提出在前兩輪對話中抽取出實體對，後續的對話透由已知實體對抽取出關係。Wei 等人 [32] 則提出 ChatIE 這種逆向的做法，在第一輪對話中先抽取關係，藉由關係來尋找相關實體。然而，多輪式生成依賴於對話 schema 的設計，且需爲每種實體和關係設計專屬模板，隨著類別增多，模板設計變得更加複雜。

通用式生成希望模型直接輸出所需的答案結構，例如 Cabot 等人 [30] 將三元組表示爲文字序列。通用式生成方法更加直觀，不受 schema 限制，且能整合不同的 Information Extraction(IE) 任務，如 Paolini 等人 [33] 將不同 IE 任務視爲自然語言的翻譯任務處理。而 UIE[1]、LasUIE[34] 模型的提出，則可以將不同的 IE 任務整合成同一種答案結構。

綜合言之，隨著大型語言模型的發展，通用式生成方法在 RE 任務上展現出色表現，即使不經過微調，僅給定良好的 Instruction，也能達到 SOTA 效能 [2]。

3 Dataset Preparation

在關係擷取的領域中，現有的資料集往往存在著來源方面的限制，也就是資料來源單一的問題。例如，NYT[1](New York Times Annotated Corpus) 來自紐約時報，DocRED 來自維基百科。此外，像是專門領域的資料集，如 GDA[2]（基因-疾病關聯語料庫）和 CDR[3]（化學物質-疾病關聯），則來自 PubMed 等生物醫學文獻庫。

爲了創建一個更具彈性和全面性的資料集，能跨越特定領域的界限，我們利用了 Common Crawl[4]，這是一個包含各種領域和文章類別的網頁文章存檔，涵蓋了多種寫作風格、觀點和主題。使我們的資料集能夠反映現實世界多樣的文本數據。

3.1 Common Crawl 數據庫前處理

Common Crawl 自 2007 年成立以來，已經累積了 17 年的網路爬蟲數據集，收集了約 2500 億筆網頁資料，橫跨 160 多種語言。其中 Common Crawl 約 1 至 2 個月會提供一次快照，更新最新的網路爬蟲數據，每次內容約 20 到 40 億個 pages 不等。而本次研究是使用 2023-50 的快照來進行處理，我們擷取了其中的 990 個 Segments，並分成 11 個 shards 進行處理。數據處理流程採用 CCNet [35] 所提出的做法，將步驟分爲: 去重、語言辨識、品質篩選等三個 pipeline 步驟。

在預處理階段，首先對每個段落進行小寫轉換，將所有數字替換爲佔位符，並消除所有 Unicode 標點符號和重音符號。

1. 去重: 去重過程通過計算每個段落的 64-bits SHA-1 雜湊值來實現。主要目標是確保網頁內容唯一性，從其他語言的網頁中移除大量的英文文本，如網頁導覽列、cookie 警告、聯絡資訊等冗餘訊息，進而降低後續語言辨識的難度。

2. 語言辨識: 我們使用 fastText[36] 作爲語言分類器，fastText 是 meta 的一個語言分類模型可以分類 294 種語言，主要在 Wikipedia、Tatoeba、SETimes 上面進行預訓練。fastText 會遍歷所有的網頁內容，並對所有語言分類打分數，我們透過該方法篩選出得分大於 0.5 的中文網頁內容。

3. 品質篩選: 最後對文件進行品質篩選。首先，我們使用 Sentence Piece tokenizer[5]把每一個網頁在句子層次做 tokenize，然後使用評分語言模型來爲每一個自然段做評分，我們使用 KenLM[6] 庫裡面的 Kneser-Ney 作爲評分語言模型，其所評的分數代表困惑度 (perplexity 簡稱 ppl) 分數，ppl 的分數越低，文本品質越高。代表其行文越流暢、越有邏輯。

總結來說，在進行品質篩選後，每一個 shards 都可以得到頭部、中間、尾部三個部

[1] https://catalog.ldc.upenn.edu/LDC2008T19

[2] https://bitbucket.org/alexwuhkucs/gda-extraction/src/master/

[3] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/

[4] https://commoncrawl.org/

[5] https://github.com/google/sentencepiece

[6] https://github.com/kpu/kenlm

份的檔案，品質狀況爲頭部 > 中間 > 尾部。我們通過該方法過濾掉了大量品質不佳的網頁資料。只留取頭部約 17% 的高品質資料作爲我們的資料來源。其中 shard 0 的頭部 (head 0)，共 26,293 筆資料，作爲測試資料集的來源，而 head 1 ~ head 10 共 260,469 筆資料作爲訓練資料集的來源。

4 Data Annotation

在這個研究中，我們選擇以人物作爲實體，並定義了 [親屬、師生、同事、其他] 作爲我們標記的 4 種關係類型。由於人工標記大量文章級別的資料既耗時又費力，我們使用 Gemini-1.0-pro(後稱 Gemini) 和 GPT-3.5-turbo(後稱 GPT) 作爲標記工具。

標記流程主要分爲四個階段 (標記流程如圖1)：

三元組生成：讓兩個模型分別處理所有文本，生成潛在的人物關係三元組。關係分類：將生成的三元組分類至預定義的四種關係類型中。交叉驗證：對兩個模型的分類結果進行比對，以解決模型分歧。資料合併：將經過交叉驗證的結果合併，形成最終的標記資料集。

爲確保標記資料的可靠性，我們採用了雙模型交叉驗證的機制。如此一來，可有效避免單一模型可能產生的偏誤，提升標記結果的準確性。因此，除非另有說明，後續的分析皆以經過雙模型驗證的測試集爲基礎。

由於訓練資料約爲測試資料的 10 倍，爲節省 API 資源，在三元組生成階段我們先讓 Gemini 過濾掉大量無關係和錯誤回覆的資料後，只針對有關係的部分，讓 GPT 進行三元組生成，如圖2，而後的關係分類、交叉驗證及合併資料流程則與測試資料集流程相同。

4.1 文章級別挑戰

現有的生成式模型多在句子級別資料上進行 few-shot 學習 [37, 2]，以實現聯合實體關係擷取。然而，我們的網頁資料篇幅較長 (平均字數 1,502 字)，若切割文章恐導致跨句實體關係的遺失。因此，我們希望模型能直接在整篇文章上進行關係擷取。

雖然我們部屬的 Gemini-1.0 和 GPT-3.5 的 context window 可以達到 16k 個 tokens。但當在長篇文本中引入 few-shot learning 範例時，過長的提示反而會影響模型性能。我們在 Gemini 上進行了小規模的測試，我們隨機取 100 筆在 zero-shot 下模型可以正常擷取關係三元組的資料，分別測試 1-shot 及 2-shots 效能，並且使用 Wadhwa 等人 [2] 的方法，人工針對這 2 筆測試資料加上 chain-of-thought(CoT) 的模板，具體的 prompt 內容詳見附錄A1。實驗結果發現，模型在給定文章級別的範例後，無論是 1-shot 或是 2-shot，其性能並未顯著提升。因此，我們決定在後續實驗中採用 zero-shot 的方式。

4.2 三元組生成

爲使大型語言模型能直接生成關係三元組，我們嘗試以通用方式引導模型。然而，在 zero-shot 設定下，同時完成命名實體識別（NER）和關係抽取（RE），並將關係類型限制在四種分類（親屬、師生、同事、其他）並不容易。因此，我們決定先讓模型聚焦於找出「有關係的人名」，暫不強求其分類關係。爲了確保模型產出符合預期格式，我們設計了 zero-shot 提示詞（如附錄A2），並採用輪對話方式，當輸出結果如果不符合我們所規定的格式，則會採用輪對話的方式，將強調格式規範的 prompt(如附錄A3) 加入到之前的對話中，並重覆以上動作直到模型回覆符合正確格式。如果連續 5 次無法依照格式回覆，我們則另外註記該筆資料。標註結果可分爲正確標註（符合格式的三元組或準確判斷無關係）和錯誤標註三類：格式錯誤（如出現四元組）、無法識別（模型回覆偏離主題）以及 API 異常（因內容不當導致模型無法回應）。

Table 1: Gemini 和 GPT 成功標記文章分析

	Gemini	GPT
正確標記	26,218	25,614
格式錯誤	29	226
無法識別	0	207
API 異常	46	246
總數	26,293	26,293

由上表1中可以看出，Gemini 在錯誤回覆的數量上，相比 GPT 還要少許多，對於格式的可控性較爲良好。再進一步統計兩個模型所正確標記的資料如表2，可以看出大部分的網頁文章是不具有人名之間的關係的，這也符合我們的認知預期。透過大型語言模型的生成標記，即可過濾掉 8 成以上的無關係網頁文章。

Table 2: Gemini 和 GPT 正確標記中，具有人名關係文章占比

	Gemini	Gpt
有關係	2,268	3,576
無關係	23,950	22,038
有關係占比	8.65%	13.93%

4.3 關係分類

在經過三元組生成的步驟後，我們可以分別得到 Gemini 和 GPT 所生成出來的三元組內容。

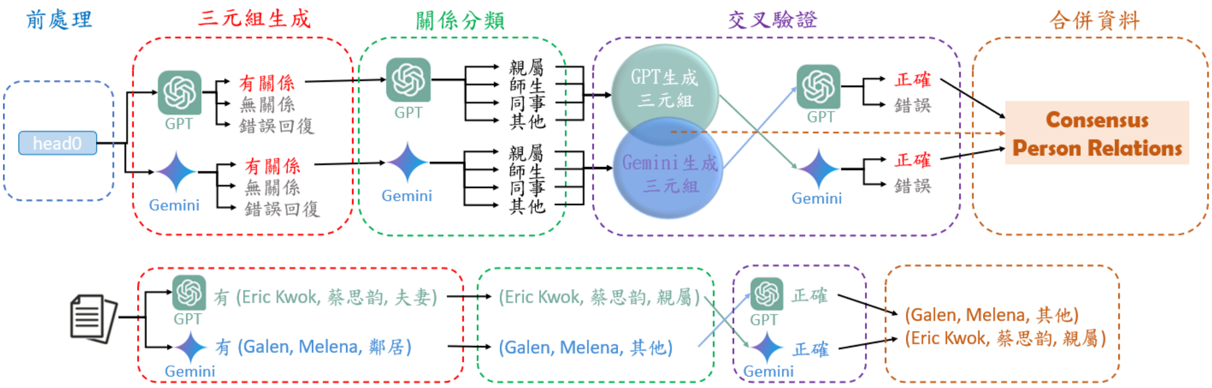


Figure 1: 測試資料集的流程 (上半) 及範例 (下半)

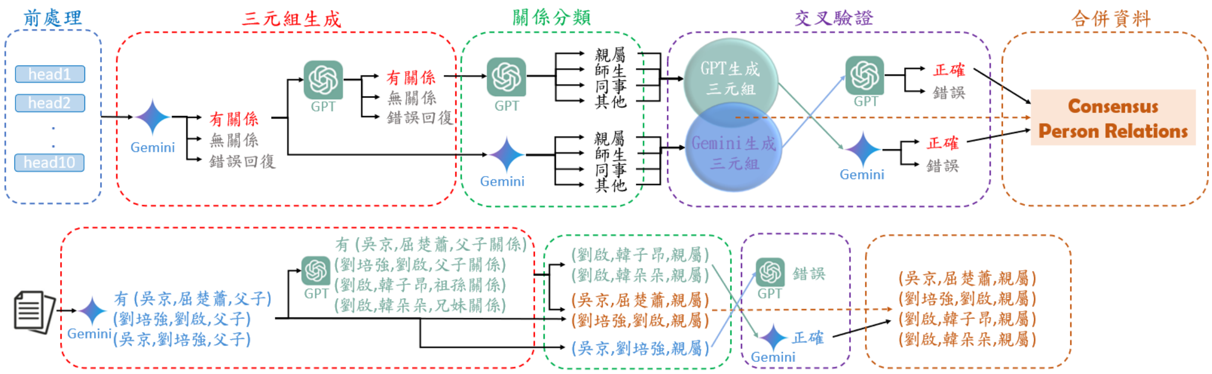


Figure 2: 訓練資料集的流程 (上半) 及範例 (下半)

然而，我們實際所生成的關係類型種類，並沒有侷限在我們所定義四種類型中。具體統計在 Gemini 所生成的三元組中，關係的種類多達 1,142 種，在 GPT 所生成的三元組中，則多達 1,825 種。而在這步驟，我們需要將這些關係種類歸類爲我所定義的四種分類 (親屬、師生、同事、其他)。

因此，我們將三元組生成關係特別取出，並將其視爲一個簡單的四元分類問題，透過 Gemini 和 GPT 以生成的方式，各自回答所生成的關係是屬於何種類別。具體指令如附錄A4所示。我們統計兩模型各自分類的數據結果，如表3。可以看出無論 Gemini 或是 GPT 都是以其他類別的種類佔多數。代表我們生成出的關係，多數被認定爲我們所指定的親屬、師生、同事之外。

Table 3: 模型標註的關係種類統計

	Gemini	GPT	佔比
親屬	135	105	8.09%
師生	61	108	5.70%
同事	194	93	9.67%
其他	752	1,519	76.54%
總關係數	1,142	1,825	100%

4.4 交叉驗證

爲驗證 LLM 標記的準確性，我們選取 Gemini 和 GPT 認爲含人物關係的 4,619 筆網頁資料聯集 (見表2)。我們對其生成的三元組進行關係分類，並統計兩模型的資料筆數和三元組數量 (見表4)。

比對時，我們採用嚴格標準: 三元組文字須完全相同，但實體順序不拘。由於標記資料可能有簡繁體混用，我們使用 OpenCC[7]將所有文字轉爲繁體後再比對。

Table 4: 有關係網頁資料筆數和生成三元組數量

	Gemini	Gpt	Inter	Union
有關係網頁數	2,268	3,576	1,225	4,619
三元組數量	6,697	8,598	1,027	14,268
三元組數量/doc	2.95	2.40	-	-

結果顯示，Gemini 和 GPT 共同認定 1,225 篇文章具人物關係，但共同認定的三元組僅 1,027 組，少於共識文章數。兩模型分別生成 6,697 和 8,595 組三元組，交集僅 1,027 組，佔比分別爲 15.34% 和 11.94%。這表明兩者生成的三元組差異較大，難以直接取得共識。

因此，我們簡化任務爲二元分類，讓兩模型

[7]https://github.com/BYVoid/OpenCC

互相評估對方生成的三元組是否正確。我們直接採用 1,027 組共識三元組，僅對無共識三元組提問 (具體指令見附錄A5)。指令中列出四種常見錯誤:A. 關係錯誤; B. 實體非人名; C. 實體僅爲稱謂; D. 兩實體相同。

交叉驗證結果見表5。儘管初始生成的三元組差異大，但兩模型對彼此生成的三元組認同度均超過 90%。進一步分析發現，當網頁含大量人名 (如演員名單、出賽名單等) 時，模型往往只取樣部分實體對並認定關係，導致標註不一致。關於模型識別人名實體的能力，我們將在章節5.1中深入探討並提出解決方案。

Table 5: 兩模型交叉驗證通過統計

生成模型	驗證模型	通過	未通過	通過比例
Gemini	GPT	5,166	504	91.11%
GPT	Gemini	7,254	317	95.81%

4.5 合併資料

我們將兩個模型原本就有共識的 1,027 組三元組，以及 Gemini 通過 Gpt 驗證的 5,166 組、GPT 通過 Gemini 驗證的 7,254 組，共 13,447 組實體關係三元組，視爲我們暫定的共識人物關係三元組。在這 13,447 組實體關係三元組，共分佈在 4,515 筆網頁文章資料中，即原本兩模型的聯集資料 4,619 筆，經過交叉驗證後排除了未通過驗證的剩餘文章數。我們統計四種關係的實際分佈狀態如表6。

Table 6: 共識人物關係三元組中，不同類型關係分佈，由於一篇文章可能含有多種關係，因此含有四種關係的文章數加總後，會大於文章總數 4,515 筆

關係類型	# 三元組	佔比	# 文章數	佔比
親屬	1,168	8.69%	642	14.22%
師生	1,192	8.86%	743	16.46%
同事	6,172	45.90%	2,196	48.64%
其他	4,915	36.55%	2,255	49.94%
總數	13,447	100%	-	-

可以看出在我們所定義的四種關係類型，在網頁文章分布是相當不平衡的，其中同事和其他關係的三元組佔了 45.90% 及 36.55%。相比表3數據，可以發現兩者佔比高的原因並不相同。我們可以看出在進行關係分類前，模型所生成的關係總類中，其他關係明顯多於另外三者，這是因爲我們將其他未定義的人與人關係，如: 朋友、同學.. 等，全部都歸類到 “其他” 關係所導致的資料不平衡。但同事關係的在種類占比中只有 9.67%，只略高於親屬和師生分別 1.58% ~ 3.97%，不過在三元組數量中卻多了超過 30% 以上。因此，我們可以推論，在我們所清洗的網頁文章中，包含同事關係的資料本身就遠超過親屬和師生的數量。

5 Data Quality Evaluation

5.1 NER 效能評估

我們在章節4.4中透過分析 Gemini 和 GPT 的三元組交集，以及通過交叉驗證的比例，發現兩模型的所標記的三元組差異很大，但卻又很高比例的認同對方所標記爲正確。透過觀察實際案例，推測是兩模型所能找到的人名實體對過少，所導致的標記不一致。

爲了驗證我們的猜想，我們使用傳統的序列標記模型，先將所有的人名找出來，我們使用的中文 SOTA-NER 模型爲中研院所開源的 ckiplab/bert-base-chinese-ner[8](後稱 CKIP)。我們將 CKIP 所標記出來的人名視爲標準答案，分別評估 Gemini、GPT，以及合併後共識人物關係三元組 (Consensus) 的 NER 效能，在 NER 效能的計算上，我們以每筆文章中的實體爲單位，計算模型預估值和正確答案的實體是否一致，兩者在使用 OpenCC 翻譯爲繁體中文後，在嚴格比對下需要完全一致才視爲相同實體，最終效能如表7。

Table 7: 將 CKIP 所標的實體視爲答案，評估模型 NER 效能

	Recall	Precision	Micro-f1
Gemini	11.53%	66.48%	19.65%
GPT	12.84%	51.92%	20.59%
Consensus	19.55%	55.47%	28.91%

由於我們最初的任務是實體關係三元組生成，而非找出所有人名實體。因此，如果人名之間並不存在關係，那該人名實體未被生成也實屬合理。但是我們可以看出即使是兩模型合併後的共識，Recall 也不到 20%。代表有超過 80% 以上的人名之間都沒有關係。因此，我們將在章節5.4中，針對 CKIP 所標記出來的人名，進行實體關係三元組的擴充實驗。

由表7還可以注意到一點，就是 Gemini 和 GPT 兩者所標記的人名實體 Precision 介於 51.92% ~ 66.48% 之間，可以看出大型語言模型找出了很多的人名實體是傳統的 SOTA-NER 所無法標記的。而這其實主要是由於三個原因所造成: 1. 嚴格比對落差: 由於我們使用嚴格比對，字符必須完全一致。但 LLM 在生成三元組時，常常會把如:“先生”、“小姐”.. 等稱謂給生成出來。進而導致與序列標記的實體不一致。2. 外文實體: 我們的資料來源包含中英文或其他外語參雜的語法。但我們的

[8] https://github.com/ckiplab/ckip-transformers

SOTA-NER 是使用 bert-base-chinese，對於英文、日文或其他外文的實體是無法辨認的。但是如 Gemini、GPT 等這些 LLM，在這些跨語言參雜的文章中找出實體是沒甚麼問題的。3. 幻覺生成: 目前幻覺問題一直存在於所有生成模型，Gemini 和 GPT 也不例外。因此，我們將在章節5.2中統計 Gemini 和 GPT 的實體幻覺比例。

5.2 幻覺評估

由於我們使用的生成式方法，可能產生不存在於文章中的實體。因此，我們以實體個數爲單位，檢查 Gemini 和 GPT 所生成出來的人名實體，是否存在原本文章中，具體統計如表8。可以發現約有 3% 左右的幻覺實體存在，而這些幻覺實體和其他實體組成實體對時，會讓含有幻覺實體的實體對比例增加到約 5% 左右，如表9所示。

Table 8: 模型所幻覺的人名實體統計，以實體個數爲單位

	Gemini	GPT	Consensus
幻覺實體數	173	480	568
總生成實體數	9,687	13,403	19,263
幻覺實體比例	1.79%	3.58%	2.95%

5.3 跨句評估

在跨句評估上，我們先將文章切割爲句子級別，使用中文常用的標點符號 [\n。；;！!？?] 和換行符號進行切割。平均每個文章會被切割爲 58.19 個句子。我們判斷實體對中的兩個人名是否有出現在相同句子，若未出現在同一個句子中，則視爲跨句子的實體對。

Table 9: 統計模型的跨句找出實體關係能力，以實體對個數爲單位

	Gemini	GPT	Consensus
含有任一幻覺實體	2.88%	6.59%	4.91%
實體對存在相同句子	63.42%	67.03%	65.28%
跨句實體對	33.70%	26.38%	29.81%

我們可以由表9中看出，模型所找出的實體關係約有 30% 左右，是屬於跨句子級別的關係。而我們進一步分析實體跨度，計算實體對間的最小間隔字數，也就是兩實體的最短距離 (Shortest Distance)。如表10所示，我們發現即便實體位置橫跨超過上千字以上，大型語言模型都還是有能力能夠找出兩者關係，這也證明了大型語言模型在跨句實體能力上傑出的表現。

Table 10: 跨句實體對中，實體對最小間隔字數

	Gemini	GPT	Consensus
跨句實體對數量	2,257	2,268	4,009
平均最小間隔字數	246	123	186
最遠的最小間隔字數	3,376	1,803	3,376

5.4 實體關係擴充

我們在章節5.1中看出，模型在 NER 任務的 Recall 效能非常不足，有超過 80% 以上的人名都未被取出。因此，我們需要對每篇的網頁文章進行實體擴充，具體流程如圖3。我們先透過 CKIP 找出文章中的所有人名實體，並將實體兩兩組成實體對，再對需擴充的實體對進行關係分類，最後合併時，去除幻覺實體關係。

由於實體對擴充時，若實體數量爲 n，實體對數量則會增爲 $\binom{n}{2}$。甚至出現如“畫廊會員名單”這種整篇文章的是人名的網頁時，最高達到 211,550 組實體對。考量到這類型的文章，並非我們最主要需要解決的類型。因此，我們設置了兩個條件來過濾掉如這種實體對過多的文章類型: 第一個條件爲人名密度，我們定義爲: CKIP entity 數量/文章字數，我們計算所有有關係的文章平均人名密度爲 0.95/100 字，我們取 2 倍的平均人名密度作爲我們第一個篩選條件。第二個條件則是實體對上限值，若實體對超過 $\binom{15}{2}$ 也就是 105 組時，我們發現很難讓模型在一次 request 中完成。

相對的，如果 CKIP 所標記的實體小於 2 個則無法組成實體對，或是組成的實體對原本就存在 Gemini 或 GPT 生成的三元組之中，我們都視爲實體對過少的文章。我們具體統計扣除實體對過多的文章類型後具體需要做實體擴充和無需做實體擴充的文章數量如表11。

Table 11: 實際需擴充佔比

	文章數	佔比
實體對過多	776	17.19%
無需擴充	1,525	33.78%
需擴充	2,214	49.03%
總數	4515	100%

在將實體兩兩組成實體對後，我們讓 Gemini 進行關係分類，並新增一個沒有關係的類別，而若模型分類爲沒有關係，則不會取用該實體。具體指令如表A6

在完成關係分類後，我們將會獲得擴充的人物關係三元組。接著，我們會在資料集中剔除實體對過多的 776 筆文章，使剩餘網頁文章數量來到 3,739 筆。隨後在去除共識人物關係中含有幻覺實體的三元組。在最終我們留下

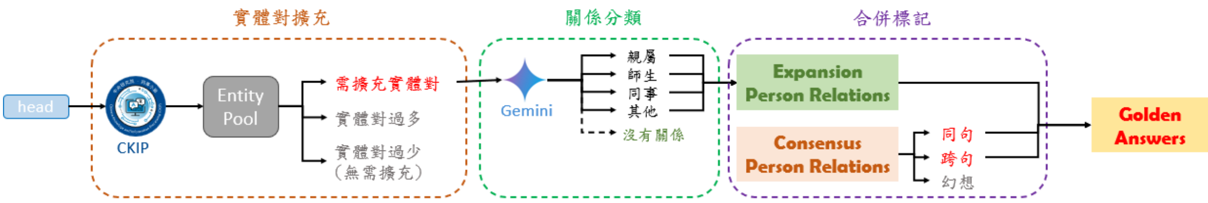


Figure 3: 實體關係三元組擴充流程

了 **3,583** 筆網頁文章，並得到共 **30,407** 組 (表12) 三元組，爲我們最終標記結果 (Golden Answers)，其中包含擴充的 21,632 組三元組，以及 Gemini、GPT 共識的 9,164 組三元組。

最後我們統計所擴充的三元組中，各個類別的分佈情形，如表12。可以看出我們所擴充的類型中，同事的類別佔有極爲懸殊的大量。符合我們在章節4.5中的預期，在眞實網路文章中，包含同事關係的文章佔比確實較高，進行實體對擴充會讓同事關係數量暴增。

Table 12: 加入擴充標記後，三元組的各類別佔比

	共識三元組	擴充三元組	Golden	佔比
親屬	783	2,231	2,967	9.76%
師生	853	157	985	3.24%
同事	3,988	18,826	22,667	74.54%
其他	3,540	418	3,788	12.46%
總數	9,164	21,632	30,407	100%

5.5 實體對評估

在得到 Golden Answers 後，我們可以回頭評估當初 Gemini 和 GPT 兩者所生成的實體關係二元組、三元組的效能。在比對二元組時，只要相同的關係實體對，我們視爲相同二元組。在表13中，我們將 Golden Answers 所取得二元組，視爲我們的 label 值。而測試 Gemini 和 GPT 在經過章節4.2三元組生成步驟後，所產生的實體對效能。

我們可以發現，由於該 label 是包含了 Gemini 和 GPT 的共識，兩者對於對方的標記大都表示認同，所以兩者的 Precision 都超過 90% 以上。但是兩者的 recall 則分別爲 14.18% 及 19.06%，可以看出 Golden Answers 所涵蓋的實體關係三元組範圍，較單一模型所能標記出的效能還要多出許多。

另外，我們也將 CKIP 所生成的實體兩兩組成對後，計算二元組效能。我們發現與 Gemini 和 GPT 不同的是，CKIP 組成的實體對 Recall 達到 82%，可以涵蓋大多數的實體對，但是由於會產生許多沒有關係的實體對，導致其 Precision 相對不佳。

Table 13: 將 Golden Answers 中實體對二元組視爲 labels，評估實體對生成效能

	Recall	Precision	Micro-f1
Gemini	14.18%	93.06%	24.61%
GPT	19.06%	93.06%	31.64%
CKIP	82.21%	41.37%	55.04%

6 Model Training

在取得由 Common Crawl 中清洗，以及透過標記流程、擴充方法得到 Golden Answers 的資料集以後，我們想驗證該資料集是否能夠在參數量較小的預訓練模型上進行微調訓練，達到通用實體關係擷取的目標。

由於考慮到設備記憶體資源有限，我們將文章截斷至最長爲 1,024 字，此外我們也把實體未出現在前 1,024 字中的 label 給去除，所以我們的測試資料筆數由 3,583 筆減少爲 **3,392** 筆，三元組數量由 30,407 組，減少爲 **21,255** 組，訓練集和驗證集也有小幅減少。

另外，爲了評估資料集規模對於模型效能的影響，我們另外進行了小規模數據的實驗。我們將 test 資料集中的 3,392 筆文章再切割成五份，進行 5-fold 交叉驗證，降低模型訓練對於資料集的偏差。

評估模型的效能時，我們將分別評估二元組效能與三元組效能。其中二元組效能，即爲章節5.5的實體對評估，我們評估最終生成的三元組中實體對部分，且不考慮實體排序，而三元組效能則需包含分類的類別也需要正確。

6.1 通用式生成

我們考慮到文章中可能夾雜中英文或日文等人名實體，我們選定多語言的 mT5-base[38] 作爲訓練的基底模型。我們實驗測試 mT5 模型是否能夠像 Gemini 和 GPT 一樣，在給定一整篇文章情形下，加上三元組生成的 prompt(如圖A2) 後，進行 full fine-tuning，讓模型直接將所有可能的實體關係三元組給生成出來。經過全微調實驗後，我們發現模型已經可以直接在三元組生成的步驟就收斂到我們所定義的 4 種關係類別，因此可省略關係分類步驟。

Table 14: 二元組效能與三元組效能

模型	方法	訓練資料	二元組效能			三元組效能		
			Recall	Precision	Micro-f1	Recall	Precision	Micro-f1
Gemini	通用式生成	NO	14.18%	93.06%	24.61%	13.39%	92.27%	23.38%
GPT	通用式生成	NO	19.06%	93.06%	31.64%	18.27%	92.42%	30.50%
mT5	通用式生成	5-fold test	18.07 %	33.26%	23.41	14.55%	26.43%	18.76
mT5	通用式生成	train data	24.93%	39.65%	30.61%	20.47%	32.12%	25.00%
CKIP+mT5	pipeline	5-fold test	71.68 %	74.94 %	73.21	64.11%	67.52	66.80
CKIP+mT5	pipeline	train data	59.50%	71.29%	64.87%	53.26%	64.29%	58.26%

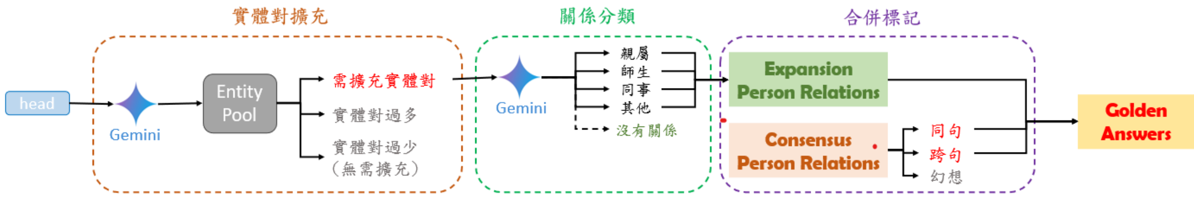


Figure 4: 使用 Gemini 取代 CKIP 實體關係三元組擴充流程

mT5 的實際效能如表14所示，在二元組效能中，我們在小規模的測試資料集內進行訓練，Micro-f1 即可達到 23.41%，已相當接近 Gemini 效能。若使用完整的訓練資料集，增加訓練資料量，即可達到 30.61% 超越 Gemini 的 24.61%，達到接近 GPT 的 31.64%。

而在三元組效能上，在小規模的測試資料集上訓練還未能超越 Gemini 和 GPT，但是在完整的訓練資料集上，即可達到 Micro-f1 25.00% 超越了 Gemini 效能的 23.38% ，但還未能達到 GPT 的效能。

6.2 pipeline

我們實驗了傳統 Pipeline 方法，使用 CKIP 的 bert-base-chinese-ner 模型進行 NER 任務。由於 Golden Answers 不包含無關的人物實體，因此未對 NER 模型進行微調。訓練時，我們將 CKIP 的人物實體兩兩配對，若該配對不在 Golden Answers 中，則將該關係標記爲 “沒有” 進行訓練。RE 任務爲 5 分類：[親屬、師生、同事、其他、沒有]，並使用 mT5 進行訓練。推論階段中，若預測關係爲 “沒有”，則刪除該實體對。由表14 可見，Pipeline 方法在小規模測試集上達到 73.21% 的 Micro-F1，遠高於 Gemini 和 GPT 效能。相較之下，在完整訓練集效能有所下降。因此對於簡單的分類問題，增加數據量不一定會有更好效果。

6.3 Gemini 實體擴充

由於我們在章節5.4進行實體關係擴充，合併產生 Golden Answer 時，部分採用了 CKIP 產生的人物實體。而 pipeline 方法的實體也來自 CKIP，使得 Golden Answer 有些許不客觀。因此，我們針對實體關係三元組擴充進行些許調整。

在測試資料集中，我們將實體對擴充所使用的模型由 CKIP 替換爲 Gemini。也就是先由 Gemini 找出人物實體，再標記關係，如圖4流程。我們給 Gemini 如表A8的指令，讓其進行 NER 任務，找出所有的人物實體，並進行兩兩配對成實體對後執行關係分類。在更新 Golden Answer 後我們重新訓練通用式生成以及 pepeline 方法實驗，並使用更新後的 Golden Answer 作爲評估指標，重新評估所有的模型。

若由 Gemini 來進行 NER 任務，會讓擴充的實體關係和大型語言模型生成的結果雷同度提高，所以 Gemini 和 GPT 的效能會因此提高。而在通用式生成的方法中，整體的 Micro-f1 效能則並未有大幅度的改變。在 pipeline 的方法中，則可以二元組 micro-f1 由 64.87% 下降至 52.39%，三元組 micro-f1 也由 58.26% 下降至 46.21%。但 pipeline 的方法還是能有優於通用式生成的整體表現。

7 Conclusion

本研究在聯合實體關係擷取領域引入了基於通用式生成語言模型的自動化標記流程，利用 Gemini、GPT-3.5 等大型語言模型，取代了人工標記的需求，提高了標記效率。且對於文章級別的標記，克服了過去因文長限制而必須截斷文本或進行證據檢索的局限性。我們利用 Common Crawl 的全網爬蟲資料庫，創建了多樣性和廣泛性的中文文章級別資料集，爲泛化的中文關係擷取研究提供了新資源。最後，我們在參數量較小的 mT5 模型中實驗證

實，該資料集所訓練出的模型，可用於泛化的眞實網路文章。總結來説，本研究爲關係擷取和命名實體識別研究提供了新的思路和資源。

References

[1] Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Unified structure generation for universal information extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[2] Somin Wadhwa, Silvio Amir, and Byron Wallace. Revisiting relation extraction in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. ACE 2005 multilingual training corpus. In *Linguistic Data Consortium*, 2006.

[4] Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA, May 6 - May 7 2004. Association for Computational Linguistics.

[5] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pages 148–163. Springer, 2010.

[6] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy, July 2019. Association for Computational Linguistics.

[7] Guoquan Dai, Xizhao Wang, Xiaoying Zou, Chao Liu, and Si Cen. Mrgat: Multi-relational graph attention network for knowledge graph completion. *Neural Networks*, 154:234–245, 2022.

[8] Linfeng Li, Peng Wang, Jun Yan, Yao Wang, Simin Li, Jinpeng Jiang, Zhe Sun, Buzhou Tang, Tsung-Hui Chang, Shenghui Wang, and Yuting Liu. Real-world data medical knowledge graph: construction and applications. *Artificial Intelligence in Medicine*, 103:101817, 2020.

[9] Jung-Jun Kim, Dong-Gyu Lee, Jialin Wu, Hong-Gyu Jung, and Seong-Whan Lee. Visual question answering based on local-scene-aware referring expression generation. *Neural Networks*, 139:158–167, 2021.

[10] Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online, July 2020. Association for Computational Linguistics.

[11] Marco Antonio Calijorne Soares and Fernando Silva Parreiras. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, 32(6):635–646, 2020.

[12] Weizhao Li, Feng Ge, Yi Cai, and Da Ren. A conversational model for eliciting new chatting topics in open-domain conversation. *Neural Networks*, 144:540–552, 2021.

[13] Yunyi Yang, Yunhao Li, and Xiaojun Quan. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14230–14238, May 2021.

[14] Ziran Li, Ning Ding, Zhiyuan Liu, Haitao Zheng, and Ying Shen. Chinese relation extraction with multi-grained information and external linguistic knowledge. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4377–4386, Florence, Italy, July 2019. Association for Computational Linguistics.

[15] Jiaqi Hou, Xin Li, Haipeng Yao, Haichun Sun, Tianle Mai, and Rongchen Zhu. Bert-based chinese relation extraction for public security. *IEEE Access*, 8:132367–132375, 2020.

[16] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li, editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the*

4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[17] Ang Sun, Ralph Grishman, and Satoshi Sekine. Semi-supervised relation extraction with large-scale word clustering. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 521–529, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[18] Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418, Florence, Italy, July 2019. Association for Computational Linguistics.

[19] Changzhi Sun, Yeyun Gong, Yuanbin Wu, Ming Gong, Daxin Jiang, Man Lan, Shiliang Sun, and Nan Duan. Joint type inference on entities and relations via graph convolutional networks. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1361–1370, Florence, Italy, July 2019. Association for Computational Linguistics.

[20] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. A novel cascade binary tagging framework for relational triple extraction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, Online, July 2020. Association for Computational Linguistics.

[21] Bowen Yu, Zhenyu Zhang, Xiaobo Shu, Yubin Wang, Tingwen Liu, Bin Wang, and Sujian Li. Joint extraction of entities and relations based on a novel decomposition strategy. In *Proc. of ECAI*, 2020.

[22] Shuai Zhang, Yongliang Shen, Zeqi Tan, Yiquan Wu, and Weiming Lu. De-bias for generative extraction in unified NER task. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 808–818, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[23] Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[24] Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. UniRE: A unified label space for entity relation extraction. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 220–231, Online, August 2021. Association for Computational Linguistics.

[25] Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. Extracting relational facts by an end-to-end neural model with copy mechanism. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[26] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32, 2019.

[27] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[29] Hongbin Ye, Ningyu Zhang, Shumin Deng, Mosha Chen, Chuanqi Tan, Fei Huang, and Huajun Chen. Contrastive triple extraction with generative transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14257–14265, 2021.

[30] Pere-Lluís Huguet Cabot and Roberto Navigli. REBEL: Relation extraction by end-to-end language generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic,

November 2021. Association for Computational Linguistics.

[31] Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. Entity-relation extraction as multi-turn question answering. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy, July 2019. Association for Computational Linguistics.

[32] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*, 2023.

[33] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

[34] Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. *Advances in Neural Information Processing Systems*, 35:15460–15475, 2022.

[35] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association.

[36] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[37] Hui Wu, Yuting He, Yidong Chen, Yu Bai, and Xiaodong Shi. Improving few-shot relation extraction through semantics-guided learning. *Neural Networks*, 169:453–461, 2024.

[38] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics.

A Appendix: Prompts for guiding LLM annotation

Table A1: few-shots 的 promt 格式及内容

1-shot
請找出以下文章中是否包含兩位具有明確姓名的人之間常見的人際關係 (例如: 親屬、師生、同事、同學...)? 且兩位關係人皆必須有明確名字，只有稱謂的不算。 若無關係直接回答:Relations: 無即可 若有請依以下格式回答: Relations: 有 (人名, 人名, 關係),(人名, 人名, 關係)... 列舉出所有關係 Explanation: 解釋原因 範例如下: TEXT: 中国计划生育观察：美国之音: 山东妇女怀孕6 月, 被强迫堕胎 下略 **805** 字**....** Relations: 有 (刘欣雯, 周国强, 夫妻) Explanation: 文章中提到刘欣雯和她的丈夫周国强在家中熟睡，可見刘欣雯與周国强爲夫妻關係 文章如下: TEXT:{document}
2-shots
請找出以下文章中是否包含兩位具有明確姓名的人之間常見的人際關係 (例如: 親屬、師生、同事、同學...)? 且兩位關係人皆必須有明確名字，只有稱謂的不算。 若無關係直接回答:Relations: 無即可 若有請依以下格式回答: Relations: 有 (人名, 人名, 關係),(人名, 人名, 關係)... 列舉出所有關係 Explanation: 解釋原因 範例如下: TEXT: 中国计划生育观察：美国之音: 山东妇女怀孕6 月, 被强迫堕胎 下略 **805** 字**....** Relations: 有 (刘欣雯, 周国强, 夫妻) Explanation: 文章中提到刘欣雯和她的丈夫周国强在家中熟睡，可見刘欣雯與周国强爲夫妻關係 TEXT: 成大材料系劉浩志團隊結合機器學習減少原子力顯微鏡量化量測誤差 下略 **2,843** 字**....** Relations: 有 (劉浩志, 阮氏芳玲, 師生),(劉浩志, 張敬萱, 師生),(劉浩志, 簡錦樹, 同事),(劉浩志, 蔡佩珍, 同事) Explanation: 文章中提及劉浩志教授與當時的博士生張敬萱與阮氏芳玲發現，可見劉浩志與阮氏芳玲爲師生關係，劉浩志與張敬萱也爲師生關係 另外文章中説到過去劉浩志教授曾與成大地科系簡錦樹教授研究嘉義布袋地底下抗砷的細菌，還有他也曾與成大醫學檢驗生物技術系蔡佩珍教授對臨床腸病毒的病毒體進行物理特性研究，所以可以得知劉浩志與簡錦樹爲同事關係，劉浩志與蔡佩珍也爲同事關係 文章如下: TEXT:{document}

Table A2: 三元組生成 Prompt，讓模型判斷內容是否具有人物關係，並限制模型的輸出格式

請找出以下文章中是否包含兩位具有明確姓名的人之間常見的人際關係 (例如: 親屬、師生、同事、同學)，且兩位關係人皆必須有明確名字，只有稱謂的不算。 若無關係直接回答: 無 即可 若有請列舉出所有關係並依格式回答: 有 (人名, 人名, 關係),(人名, 人名, 關係) 文章如下: {document}

Table A3: 強調回覆格式 Prompt，會在收到模型錯誤回覆時加入多輪對話中，以增加模型格式的控制。

請務必依照規定格式回答，若無關係直接回答: 無，若有請依 2 個人名實體和 1 個關係格式回答: 有 (人名, 人名, 關係),(人名, 人名, 關係)

Table A4: 關係分類 Prompt，該指令設計爲簡單的四元分類問題

請將以下的關係進行分類成 [師生關係、同事關係、親屬關係、其他關係]4 種類別 如果是師生關係: 請回答 師生 如果是同事關係: 請回答 同事 如果是親屬關係: 請回答 親屬 如果是其他關係: 請回答 其他 關係: {博士生指導教授與博士生} 請問是 師生、同事、親屬、其他 哪一個?

Table A5: 交叉詢問 Prompt，設計成是非題的題組，簡化任務難度，且透過一次尋問多個三元組的方式，減少 request 次數，以節省資源

分析以下文章中的人名關係三元組 (人名, 人名, 關係)。找出親屬、師生、同事等三種關係，其餘標爲其他，即類別：[親屬、師生、同事、其他]。 文章如下: {document} 關係如下: {1.(邰智源, 林柏昇, 其他) 2.(邰智源, 泱泱, 其他) 3.(邰智源, 溫妮, 其他)} 請問以上 {3} 個人名關係三元組，分別是正確或錯誤? 以下 4 種情形視爲錯誤: A. 關係錯誤，例如:(蔣中正, 蔣經國, 同事)，正確關係應爲 (蔣中正, 蔣經國, 親屬)。 B. 人名實體並非人的姓名，例如:(習近平, 共產黨, 同事)，因爲``共產黨" 並非人的姓名。 C. 人名實體沒有明確人名或是綽號，只有稱謂，例如:(湯姆·克魯斯, 妻子, 親屬)，並沒有給出妻子姓名。 D. 兩個人名相同，例如:(徐志摩, 徐志摩, 其他)，兩個人名相同即視爲錯誤。 請依格式回答: {1. 正確/錯誤 2. 正確/錯誤 3. 正確/錯誤}

Table A6: 實體對關係分類 Prompt，讓模型判斷每組人物實體對的關係

根據以下文章，判斷文中每組人名實體對的人物關係。人物關係分爲親屬、師生、同事、其他、沒有，共 5 種類型。 人名實體對：{1.(丁淑君, 林慶芳) 2.(林慶芳, 梁作磊) 3.(丁淑君, 梁作磊)} 文章如下: {document} 回答格式：{1. 親屬/師生/同事/其他/沒有 2. 親屬/師生/同事/其他/沒有 3. 親屬/師生/同事/其他/沒有 }

Table A7: RE 任務 Prompt，設計成 5 分類問題，每次只詢問一組實體對，避免增加問題複雜度

根據以下文章，找出 {person1} 與 {person2} 中之間的關係。關係分爲: 親屬關係、師生關係、同事關係、其他關係、沒有關係，共 5 種。 文章如下: {document}

Table A8: NER 任務 Prompt，讓模型找出所有的人名實體

請找出以下文章中所有的人名，並依格式回答: (人名 1, 人名 2, 人名 3...)，若文章中沒有具體人名，則回答: 無 文章如下: {document}